UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO BIOLOGIA ANIMAL

Peopling of Greek Islands: Understanding the Bronze Age transition with Ancient Genomes

Francisco Filipe Coroado Santos

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:

Professor Doutor Vítor Sousa

2019

Acknowledgments

First, I would like to thank my supervisor Vítor Sousa for this opportunity and all the help, knowledge and guidance during the last year. He is a true inspiration. I would like to thank my colleagues and friends from the Evolutionary Genetics group in cE3c for the support they gave me during this year.

I would like to thank my colleagues with whom I have worked with for the past year, they are my supervisor, Vitor Sousa, Christina Papargeogopoulou, Anna-Sapfo Malaspinas, Florian Clemente, Samuel Neuenschwander, Oscar Lao, Olga Dolgova, Martina Unterländer and J. Víctor Moreno.

I would like to thank Dr. Georgia Mentessidi-Karamitrou Ephor Emerita and Dimitra Theodorou, archaeologist, Ephorate of Antiquities of Kozani, from the Greek Ministry of Culture for kindly allowing me to use the pictures from Logkas samples.

I would also like to thank my grandparents Amélia, Isaías, Madalena and Jorge, my parents Jorge and Sofia, my sister Maria, my girlfriend Sara and all my family and friends for all the love and support.

Abstract

The field of archeology has been revolutionized by genomic studies analyzing ancient DNA from human archeologic findings. Ancient genomes from humans allowed us to reconstruct the past demographic history, revealing major migration events. However, ancient DNA has several preservation problems which usually compromise the sequencing quality of the samples. These problems are more prevalent in warmer regions, like Greece, resulting in a low depth of coverage. Therefore, there is a need to develop bioinformatic tools and pipelines to analyze such low coverage data. On my thesis I focused on the dispersal and spread of farming from Asia into Greece from the Neolithic (10,500-5,000 BP) and the transition to the Bronze age (5,000-2,500 BP). The Neolithic was first characterized by the appearance of farming settlements and later by the appearance of pottery. But it was not until the Bronze Age that the first civilizations started to appear in Europe. According to archeological data, some of the first civilizations that arose in Europe were the Minoans (Crete), the Mycenaeans (mainland Greece) and the Cycladics from the Cycladic Islands. Genomic data can be used to infer population structure and past effective sizes of populations. This can be done by detecting regions of the genome for which an individual is homozygous for successive variable sites (Runs of Homozygosity-ROH). Individuals with more and longer ROH reflect either a low population effective size or consanguinity.

In this thesis, we analyzed ancient DNA samples to investigate if the transition from the Neolithic to the Bronze Age in Greece was associated with movement of people (and hence with a genetic component) or if it happened as a cultural diffusion process without a major genetic shift. By comparing ancient samples with other ancient and present-day people we inferred how population structure changed through time. Finally, we estimated ROH to test if there were differences in the inbreeding levels of samples from islands and mainland Greece, which could reflect different effective sizes. Because of the lack of tools to accurately call ROH in low coverage data, we developed a heuristic approach to detected small stretches of homozygous sites using genotype likelihoods. We also developed a pipeline to merge modern and ancient genomes with low depth of coverage by sampling one read per site per individual, and made all scripts publicly available on github.

We analyzed six ancient genomes of newly sequenced Greek samples, belonging to three different cultures, from the beginning of the Bronze Age. We then merged them with already published ancient and modern genomic data from other relevant populations, such as Anatolians, Eurasian hunter-gatherers, populations from the steppes and Caucasus. Our results indicate that compared to Neolithic Greeks, during the Bronze Age there is an increase in the proportions of ‘a northern’ European Hunter-Gather related component and an ‘eastern’ Caucasus/Iranian-related component in Greece. This suggests that the transition to Bronze Age was associated with a genetic shift, one related to people from the Caucasus and later one related to European Hunter-Gatherers. Furthermore, we found that modern day Greeks and Cretans share the same ancestry as two of our newly sequenced middle Bronze Age individuals from northern Greece.

By applying our ROH method on chromosome 21 we identified small stretches of homozygous sites in the six newly sequenced individuals. However, we only detected small ROH (< 0.31 Mb) and further validation of the method is required.

Key words: *ancient DNA, Admixture, Runs of Homozygosity, Neolithic, Bronze Age*

Resumo Estendido

O ramo da arqueologia tem sido, ao longo dos anos, alvo de uma revolução proporcionada pelos avanços de estudos genómicos e desenvolvimento de ferramentas computacionais. A possibilidade de extrair ADN de restos arqueológicos datados de há milhares de anos (*ancient DNA*), permite que se possa reconstruir com bastante detalhe o passado demográfico das populações, como migrações ou eventos de fluxo de genes. No entanto, extrair DNA deste tipo de materiais apresenta diversas dificuldades: i) alguns métodos de extração são bastante destrutivos, pelo que técnicas menos invasivas têm sido desenvolvidas (um exemplo é extrair ADN do osso temporal onde há maiores concentrações de ADN não contaminado); ii) com o decorrer do tempo ocorre degradação do ADN que fica mais fragmentado e danificado, ocorrendo um processo denominado por deaminação que resulta na troca de purinas (G → A) e pirimidinas (C → T); iii) as amostras contêm sempre contaminação por outras fontes de ADN (bactérias, fungos ou até das pessoas que manuseiam a amostra); iv) as condições ambientais também têm muitas implicações na preservação do ADN, por exemplo zonas com temperaturas baixas conservam melhor o material genético, e ainda é complicado obter amostras de qualidade em zonas mais quentes. Por estes motivos é importante desenvolver métodos computacionais e bioinformáticos que permitam analisar *ancient DNA* tendo estes aspetos em conta, como o facto de grande parte das amostras apresentarem uma baixa cobertura (*depth of coverage*) ou os padrões de deaminação poderem causar enviesamentos.

Com o aumento de genomas humanos antigos disponíveis vários autores têm tentando perceber a expansão da agricultura na Europa, que começou no Neolítico sensivelmente há 10,500 anos no Crescente Fértil e que coincidiu com o surgimento dos primeiros povos sedentários, que se estabeleceram no Neolítico. No entanto, foi apenas na idade do Bronze (5,500 – 2,500 BCE) que as primeiras civilizações começaram a emergir na Europa, na zona do mar Egeu. Destacam-se três civilizações: os Micénicos, que pertenciam à cultura Heládica que se estendia de norte a sul da atual Grécia, os Minoicos na ilha de Creta e os Cicládicos, nas ilhas Cíclades. O facto de viverem em grandes centros urbanos e do registo arqueológico sugerir que existia contacto entre os povos do Egeu, indica que este período foi associado a um aumento do efetivo populacional. Ao nível genómico, espera-se que indivíduos pertencentes a populações com menor efetivo populacional ou descendentes de cruzamentos de parentes próximos apresentem longas regiões do genoma onde são homozigóticos para posições variáveis sucessivas (*Runs of Homozigosity - ROH*).

Nesta tese investigámos se na Grécia a transição para a Idade do Bronze esteve associada a movimentos de pessoas (e como tal variações de componentes genéticas), ou se ocorreu como parte de um processo de difusão cultural (sem grandes alterações de componentes genéticas). Ao comparar genomas do Neolítico, Idade do Bronze e de populações atuais, caracterizámos como é que a estrutura populacional se alterou ao longo do tempo, permitindo perceber a relação entre os gregos do continente e das ilhas. Com as nossas análises conseguimos ainda avaliar qual a relação entre os habitantes atuais (quer do continente quer das ilhas), com as populações que habitavam as mesmas regiões durante o Neolítico e a Idade do Bronze. Por fim, calculámos *ROH* de forma a testar se o provável aumento populacional associado a centros urbanos e aumento de comunicação entre populações do mar Egeu se reflete ao nível do genoma, no número e tamanho das ROH. Dado a falta de métodos computacionais para lidar com a limitação de trabalhar com baixa cobertura (low depth of coverage), característica de *ancient DNA*, desenvolvemos um método heurístico para calcular ROH em genomas de indivíduos com baixa coverage.

Analisámos genomas recentemente sequenciados de amostras da idade do Bronze (datadas de 2,890-1,831 BCE) que correspondem às amostras mais antigas pertencentes a cada uma das três civilizações Gregas. As amostras das ilhas Ciclades, Creta e dos Micénicos da região da Peloponésia correspondem ao início da Idade do Bronze, enquanto que as amostras dos Micénicos do norte da Grécia são do Meio da Idade do Bronze. Estes seis indivíduos foram sequenciados pelo método de *shotgun whole-genome sequence* (WGS). De destacar que as amostras Cicládicas são as primeiras, até à data, a terem sido sequenciadas em indivíduos pertencentes à civilização Cicládica. Os dados genómicos destes indivíduos foram comparados com dados de *SNP array* (Human Origin Affymetrix array que tem SNPs neutrais e que por isso a frequência alélica nesses SNPs é apenas afectada pela demografia e não por seleção) já publicados de humanos modernos e antigos, desde o Neolítico à idade do Bronze, incluindo Anatólios (atual Turquia), caçadores recolectores (da Escandinávia, Rússia e Europa Oeste), Micénicos e Cretenses do final da Idade do Bronze e povos do Cáucaso, Irão e regiões circundantes, entre outros. Para analisar conjuntamente dados de *SNP array* e de *WGS* tendo em conta a baixa cobertura (*low depth of coverage*) desenvolvemos um script que amostra uma *read* ao acaso de cada posição variável no genoma para cada individuo.

Após juntar as nossas amostras aos dados existentes obtivemos um total de 2,399 indivíduos e 165,447 SNPs. Utilizámos o software *ADMIXTURE* para detetar clusters e inferir a proporção de cada cluster em cada indivíduo. Esta análise permitiu perceber se a transição do Neolítico para a idade do Bronze está associada a alterações genéticas nas proporções de diferentes clusters e se existem diferenças entre continente e ilhas. Permitiu-nos também perceber a relação entre Gregos e Cretenses modernos com as populações que habitaram a mesma região no Neolítico e na Idade do Bronze.

Os nossos resultados indicam que os agricultores Gregos e Anatólios do Neolítico eram bastante homogéneos, com quase 100% de um dos clusters. No entanto, nos Gregos da Idade do Bronze os indivíduos começam a ter proporções de outros clusters, o que indica que têm uma ancestralidade diferente dos Gregos do Neolítico. Os nossos resultados sugerem que ocorreram migrações na zona da Grécia durante a idade do Bronze. De destacar o facto de estimarmos o aumento da proporção de duas componentes genéticas na Grécia durante a Idade do Bronze: uma associada a populações relacionadas com caçadores recolectores europeus e outra associada a populações vindas de leste, nomeadamente do Cáucaso e Irão.

Nos seis indivíduos sequenciados representantes das três civilizações Gregas foi possível distinguir dois grupos: (i) Minóicos, Cicládicos e Micénicos, do sul da Grécia (Peloponésia), que têm duas componentes genéticas associadas aos povos agrícolas do Neolítico e povos do Cáucaso e Irão; (ii) amostras do norte da Grécia, que para além das duas componentes referidas apresentam uma terceira componente relacionada com os caçadores recolectores europeus e povos das estepes. A elevada proporção da componente relacionada com os caçadores-recolectores nos indivíduos do Norte da Grécia, quando comparado aos restantes, pode-se dever a: (i) um maior efetivo populacional, o que mantém a diversidade genética ancestral; (ii) estrutura populacional resultante de um contacto reduzido das populações do norte da Grécia com as do sul e ilhas; (iii) fluxo genético durante um período mais prolongado com populações de caçadores-recolectores que contêm uma elevada proporção dessa componente. Verificámos ainda que os Gregos e Cretenses modernos apresentam as componentes em proporções semelhantes àquelas que nós encontrámos em amostras da Idade do Bronze no norte da Grécia.

De forma a averiguar se o possível aumento dos centros urbanos e comunicação entre populações do mar Egeu se reflete nos níveis de *inbreeding*, aplicámos o método que desevolvemos para detetar ROHs nos genomas das seis amostras da Idade do Bronze. O método que desenvolvemos, ainda que preliminar,

permite detectar ROH para dados com coverage reduzida utilizando *genotype likelihoods*. Para tal definimos uma probabilidade limiar (P) para considerar um site homozigótico (ex: $P > 0.8$). O método é flexível e permite considerar que um determinado site, com por exemplo uma probabilidade de ser homozigótico *P* entre 0.5 e 0.8, seja incluído numa ROH no caso de nas posições circundantes a probabilidade de ser homozigótico seja acima do limiar ($P > 0.8$). Aplicando este método ao cromossoma 21, as ROHs detectadas são pequenas (a maior com 0.31 Mb). Estas análises para detetar ROH foram limitadas ao cromossoma 21 devido a limitações de tempo. Utilizámos também um método padrão (*PLINK*) mas com esse método não foi possível detetar ROH. No futuro, seria importante testar o *PLINK* e o nosso método para os restantes cromossomas.

Os scripts que desenvolvemos para analisar dados com baixa cobertura (*low depth of coverage*) estão disponíveis no github, incluindo o script para juntar dados de WGS com SNP array amostrando uma read ao acaso, assim como o método para detectar ROHs com base em *genotype likelihoods*.

Palavras-Chave: *ancient DNA, Admixture, Runs of Homozygosity, Neolítico, Idade do Bronze*

Contents

List of Figures and Tables

Acronyms

BA – Bronze Age

BAM - Binary Alignment Map

BP – years before present, being present the year 1950

BCE – before current epoch

CHG – Caucasus Hunter-Gatherers

ChL - Chalcolithic

DNA - Deoxyribonucleic acid

EBA – Early Bronze Age

EHG – Eastern Hunter-Gatherers

EMBA – Early to Middle Bronze Age

HG – Hunter-Gatherers

IA – Iron Age

LD – Linkage Disequilibrium

MAF – Minor Allele Frequency

Mb – Mega base pairs

MBA – Middle Bronze Age

MLBA – Middle to Late Bronze Age

MNChL – Middle Neolithic to Chalcolithic

N – Neolithic

ROH(s) – Run(s) of Homozygosity

SHG – Scandinavian Hunter-Gatherers

SNP(s) – Single Nucleotide Polymorphism site(s)

VCF – Variant Call Format

WHG – Western Hunter-Gatherers

1. Introduction

The field of ancient DNA has seen tremendous developments in the last decade, especially due to new high throughput sequencing technologies. It is now possible to sequence whole genomes from ancient DNA fragments, allowing to test hypothesis of human population history and evolution made by anthropologists and archaeologists (Skoglund and Mathieson 2018). However, extracting DNA from ancient remains poses challenges: 1) The process to retrieve DNA implies destruction of part of the sample (e.g. skull, teeth, bones, etc.). This has improved in recent years as researchers developed less destructive extraction methodologies relying on extracting DNA from portions with high amounts of endogenous DNA, such as the petrous bone (Pinhasi et al. 2015); 2) ancient DNA is highly fragmented and damaged, with common miscoding lesions in ancient DNA (deamination) causing a change in purines (A → G or G → A) and in pyrimidines (C → T or T → C) (Binladen et al. 2006; Dabney, Meyer, and Pääbo 2013); 3) The analysis of ancient human DNA is associated with a high risk of contamination from other sources of DNA (e.g., bacterial, fungal, etc.) or even from the person who handles the samples, not only in laboratory, but also in the archaeological site; 4) Environmental conditions like humidity, pH, salinity and temperature influence the degradation of DNA over time (Dabney, Meyer, and Pääbo 2013). The DNA of ancient samples from Siberia and other colder regions is better preserved than in samples from warmer regions because low temperatures and high salt concentration increase the longevity of DNA molecules (Willerslev and Cooper 2005). Hence, obtaining ancient DNA from warm regions, for instance from tropical latitudes, poses several challenges. All these issues need to be considered when deciding if samples are preserved well enough to be worth sequencing, and when deciding on the methods and approaches to analyze them. With the sequencing revolution and the possibility of getting good quality genomes at a lower cost, genomic studies on ancient DNA from human archaeologic findings is becoming more common (Figure 1.1). This allowed to reconstruct the genetic history and evolution of modern humans in more detail (Skoglund and Mathieson 2018). For instance, the sequencing of a Neanderthal genome resulted in the discovery that most Eurasians individuals have approximately 2-4% of Neanderthal DNA (Green et al. 2010); and modern day Melanesians have approximately 4-6% of Denisovan DNA (Reich et al. 2010). Also, the sequencing of Siberian modern humans showed that the history of Siberia was associated with at least three migration waves that resulted in population replacement (Sikora et al. 2019).

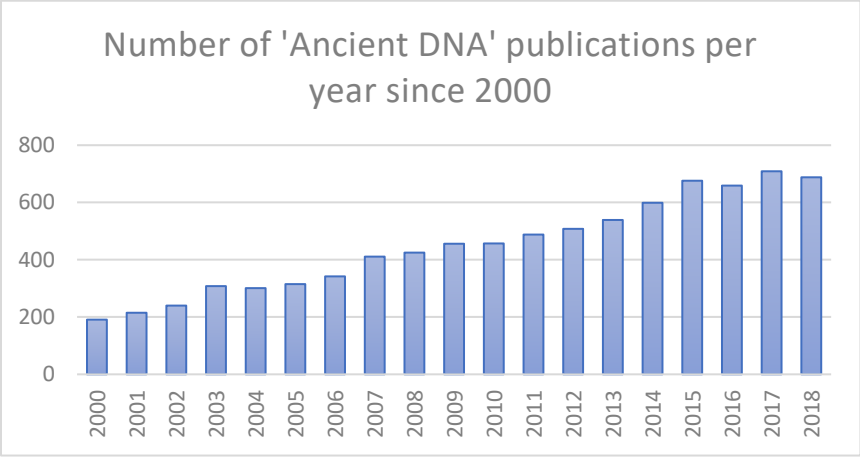

Figure 1.1 – Number of publications related to Ancient DNA, per year, have been increased over the last two decades (data since 2000, obtained in Web of Science in 16/10/19, using the keyword "ancient DNA").

Studying the DNA from ancient individuals has been used to understand the agricultural revolution, that started around 10,500 BP in the Fertile Crescent (Mathieson and Mathieson 2018). The process in which farming was introduced in Europe from Anatolia during the Neolithic (10,500-5,000 BP) is complex. Two main hypotheses have been suggested to describe the process. One suggests it could have originated from cultural diffusion without significant gene flow between the farmers and the local Hunter-Gatherers (HG), hence suggesting that cultural knowledge transfer was the main process. The other is a demic expansion hypothesis that suggests that peoples' movement and migration was the most important process, with Hunter-Gatherers populations being replaced by farmers (Mathieson and Mathieson 2018; Hofmanová et al. 2016; Fort 2015). Genetic studies indicate that the diffusion of agriculture in Europe is more complex than these two extreme hypotheses, as mixing between HG and farmer populations likely occurred (González-Fortes et al. 2017). Ancient DNA studies suggest that Anatolia Neolithic migrants did not completely replace HG, as by 4,500 BP almost all Europeans are estimated to share genes from HG and Anatolia farmers (Skoglund and Mathieson 2018).

After the Neolithic, during the Bronze Age (5,000-2,500 BP), the first civilizations characterized by monumental palaces started flourishing in Europe, especially in Greece. In South East Europe, there were intense commercial networks surrounding the regions of the Aegean Sea (Lazaridis et al. 2017). Archeological data suggest that there were at least three civilizations in Greece, which are among the first in Europe: the Minoans, who were from Crete; the Mycenaeans (who were part of the Helladic culture), from mainland Greece (Lazaridis et al. 2017) and the Cycladic Culture from the Cycladic Islands (Broodbank, 2000). Studying the Early Bronze age in Greece is important because of its geographic location, namely the proximity with the origin of agriculture in the middle east; and because it allows us to understand if the spread of cultural innovations was different between islands and inland; and whether it involved movement of people and admixture or cultural transmission. *Ancient DNA* studies suggest that the Yamnaya pastoralist culture, represented by populations from the Eurasian steppes, likely migrated into Europe during the Early Bronze Age. This migration event is likely responsible for the observed major genetic shift seen across time as we go from older to more recent ancient DNA samples (Haak et al. 2015), especially in the regions between the Volga and Rhine rivers, which show a close genetic affinity between Yamnaya and the Corded Ware cultures (Juras et al. 2018). The Yamnaya influence had a very important impact on many levels: 1) modern day Europeans harbor a genetic legacy from Yamnaya ancestry (Juras et al. 2018); 2) although the Yamnaya were not the first people to domesticate the horse (which is widely attributed to the Botai (Gaunitz et al. 2018)), they were the first to domesticate and bring to Europe the ancestral of present-day horses (de Barros Damgaard et al. 2018) (however, we cannot reject the hypothesis of the occurrence of horse domestication in the Iberian peninsula independently of the Yamnaya (Orlando 2019)); 3) Yamnaya are the most likely source for the introduction of the Proto Indo-European language into Europe ('steppe hypothesis'), which is postulated to be the origin of (almost) all languages spoken in Europe and Asia (Haak et al. 2015); 4) The steppe populations, and especially the Yamnaya, had an increased allele frequency for the lactose persistent (LP) mutation (in the SNP position rs4988235) (Allentoft et al. 2015). That mutation allows people to degrade lactose during adulthood (Gerbault et al. 2011) and, even though Neolithic pottery shows remains of dairy fat (Copley et al. 2003), the LP mutation was absent in Neolithic Europeans (suggesting animals were milked before the LP mutation increased in frequency (Burger et al. 2007)). It was only in Bronze Age that Europeans started to have an increase of the derived allele frequency (~5%) and the region with the most Yamnaya influence, the Corded Ware, and Scandinavia showed the highest frequency among all Europeans (Allentoft et al. 2015).

Present-day genomes of Europeans can be modeled as having ancestry derived from three main sources, that could be seen as three ancestral populations: (a) the Neolithic European Farmers; (b) Steppe-related populations (like the Yamnaya); and (c) from populations from the Caucasus (Lazaridis I, Patterson N, Mittnik A, 2014). In order to better understand how these three sources appeared and spread in Europe it is important to understand the Neolithic to Bronze Age transition. Since Greece is one of the oldest BA transitions in Europe and because it is geographically close to the Caucasus, analyzing BA Greek samples can shed light on the movement of people from Caucasus and Yamnaya, and how those influenced European genomic diversity. Lazaridis *et al.* (2017) performed the first genomic analysis on ancient Minoans and Mycenaeans. Their results revealed that Minoans had an Anatolia Neolithic 'farmer' and an 'eastern' Caucasus-related ancestry while Mycenaeans from BA Greece, beside those two also have a northern ancestry (as they can be modelled as a mixture of Minoans and north ancestry sources like: Steppe populations and Easter Hunter-Gatherers, among others) (Lazaridis et al. 2017). However, Lazaridis et al. (2017) study had some limitations: (i) the samples were sequenced using a target capture technique with a very low depth of coverage (median of 0.87x), and (ii) they lack samples from the Cycladic culture. The fact that archeological evidence supports intense contact and exchange between different civilizations in the Aegean (Lazaridis et al. 2017; Biehl P. 2008), raises the question of whether this trading network was also associated with gene flow between populations. Furthermore, if gene flow was limited, comparing mainland with island populations could elucidate if there were differences in the population effective sizes. In more isolated populations (for instance, in islands) we would expect smaller effective sizes and higher chances of sampling inbred individuals with higher relatedness between each other and less genetic variation within the population. In contrast, if gene flow occurred, we would expect to find evidence of admixture between mainland and island populations.

We can use genomic data to understand whether a group of individuals are related or if they belong to a population with higher or lower effective size and hence with higher inbreeding. A way to assess the level of inbreeding of an individual is to look at the length of contiguous regions where an individual is homozygous across variable sites (Runs of Homozygosity - ROH) (Ceballos et al. 2018). Longer homozygous segments may reveal that a given individual is the product of recent consanguinity mating or cultural practices where individuals tend to mate with close relatives, increasing homozygosity, and hence tending to have not only a higher number but also longer ROHs (McQuillan et al. 2008). At the population level, the number and length of ROHs depend on the demographic history and hence ROHs can be used to reconstruct the population's demographic history. For instance, simple models show that populations that result from admixture between differentiated populations are expected to have fewer and shorter ROHs, when compared to isolated populations. Also, populations that go through bottlenecks are expected to have more and longer ROHs than a constant size population. A higher number and length of ROHs can increase the frequency of recessive deleterious mutations, which can have harmful effects and increase the prevalence of genetic diseases (for example, Tay-Sachs syndrome) (Ceballos et al. 2018). To detect Runs of Homozygosity there are two main approaches: observational and model based. The observational approach is implemented in *PLINK* (Chang et al. 2015) and it is the standard method. After filtering a genome for SNPs, it uses a sliding-window of a given size to detect regions of the genome from which the individual is homozygous (Ceballos et al. 2018). Because SNPs are variable sites within a population, we would expect that an individual with closely related parents (or belonging to a population with low effective size) would have multiple SNPs for which it is homozygous. The observational approach of *PLINK* outperforms the computationally expensive model-based approaches (Ceballos et al. 2018). ROHs can be classified into three categories: very short ones (around 100 Kb) that reflect LD patterns, intermediate

(around 2Mb) that result from background relatedness owing to genetic drift, and long (> 2Mb) that are due to parental relatedness (Ceballos et al. 2018).

In my thesis I will use newly sequenced samples from the three Greek civilizations described above. They are dated from Early and Middle Bronze Age and are the first whole genome sequenced samples to date and to our knowledge, from Early Bronze Age in Greece. Previously published Mycenaeans and Minoans samples date from the Late Bronze age, so by having Early and Middle Bronze samples from south and north mainland Greece and the islands, we fill a gap in time from which there were no samples. We analyze these genomes to understand the transition from Neolithic to Bronze age in Greece.

2. Objectives

The general goal of my thesis is to provide a better understanding about the transition from Neolithic to Bronze Age and evaluate how it impacted the genomic ancestry of modern populations. For that, I used newly sequenced ancient DNA samples from the Early Bronze Age (EBA) in Greece to characterize the population structure and levels of inbreeding in southeast Europe during the transition from Neolithic into Bronze Age. Regarding the population structure, the specific objectives are:

1) To estimate the relationship of EBA and MBA Greek samples with other Eurasian samples from Neolithic, Bronze Age and present-day Europeans;
2) To test if the estimated genetic structure reflects the cultural divisions from the three Bronze Age civilizations that emerged in Greece, by comparing the Early Bronze age Helladic, Cycladic, Minoan and Mycenean samples.

Regarding the levels of inbreeding, the specific aims are:

1) To develop a heuristic approach accounting for genotype call uncertainty to infer levels of inbreeding based on Runs of Homozygosity from low coverage data (1.00-5.00x);
2) To compare levels of inbreeding in mainland and island populations from EBA in Greece to test whether the increase of communication in Bronze Age Aegean reflects higher effective size on islands.

3. Methodology

3.1 Sampling and NGS

To study the transition from Neolithic to Bronze Age we sequenced the oldest Bronze Age individuals sampled in Greece to date (Figure 3.1 and Table 3.1). Our collaborators sampled and sequenced the whole genome of six individuals from the petrous bone (Figure 3.1) to a depth of coverage between 2.6 and 4.90x (Table 3.1). The sample preparation was done as described in (Hofmanová et al. 2016) and the DNA extracts were converted into sequencing libraries following (Kircher, Sawyer, and Meyer 2012) protocol. The DNA extraction and library preparation were done in the facility of the Palaeogenetics Group at the Johannes Gutenberg-University, Mainz. The sequencing was done using an Illumina HiSeq3000 at the Lausanne Genomics Technologies Facility. After sequencing, each read was trimmed by 5bp at both ends, which reduced the overall error rate to 0.24%. There was an increased C to T and G to A substitutions at the end of the read which is consistent with the damage patterns in ancient DNA. Reads shorter than 30bp were then removed. The reads were aligned to the human genome 37.1 and alignments with a quality score below 30 were discarded. Our samples have between 0.01 and 1.49% of contamination (95% credible interval). These bioinformatic pre-processing of the data were done by our collaborators from the group of Prof. Anna-Sapfo Malaspinas at the University of Lausanne. Based on the archeological context, these individuals could be attributed to the civilizations that arose in the Bronze Age Greece: Mycenaeans, Minoans and Cycladic's. Samples from Logkas (Figure 3.2) are not completely defined as Mycenaeans, but they are considered Middle Bronze Age Early Helladic individuals (broad term used to characterize Bronze Age Greece culture, from which Mycenaeans belong), from Northern Greece. Mik15 is an Early Bronze Age individual belonging to Early Mycenaean culture. Pta08 is an Early Bronze Age Minoan from Crete, and Kou01 and Kou03 (Early Bronze Age) are the first and only samples from the Cycladic civilization.

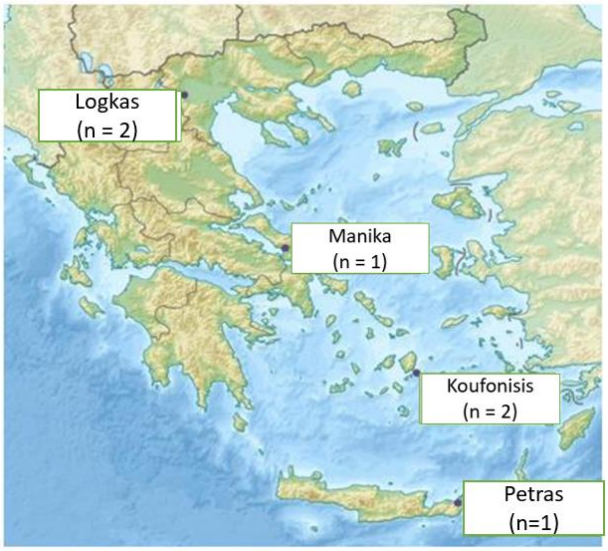

Figure 3.1 Geographical location of our samples in Greece. We have two samples from Logkas (Log02 and Log04) and represent the Helladic culture. Manika has one sample (Mik15) and represents Early Mycenaeans, also part of the Helladic cilture. Koufonisis has two samples (Kou01 and Kou03) and represent Early Cycladic's. Petras has one sample (Pta08) and represents Early Minoans.

Table 3.1 Details about the samples label, culture, age (radiocarbon calibrated dated in Curt-EngelhornZentrum Mannheim), method of sequencing, coverage and sex. The sample with higher coverage is Log04 with 4.90x and the one with less coverage is Kou01 with 2.61X.

Site	Sample	Culture	Age (BCE)	Method	Coverage	Sex
Koufonisi	Kou01	Early Cycaldic	2464-2349	Shotgun	2.61	XY
Koufonisi	Kou03	Early Cycladic	2832-2578	Shotgun	2.81	XX
Logkas	Log02	Middle Helladic	1924-1831	Shotgun	4.34	XX
Logkas	Log04	Middle Helladic	2007-1915	Shotgun	4.90	XX
Manika	Mik15	Early Mycenaean	2890-2764	Shotgun	3.54	XX
Petras	Pta08	Early Minoan	2849-2621	Shotgun	4.00	XY

Figure 3.2 Picture of our two Logkas samples - Log02 (left) and Log04 (right). Credit to Dr. Georgia Mentessidi-Karamitrou Ephor Emerita, Greek Ministry of Culture and Dimitra Theodorou, archaeologist, Ephorate of Antiquities of Kozani, Greek Ministry of Culture

3.2 Population Structure – Individual based admixture estimates

3.2.1 Estimating admixture: Principles and assumptions

There are several methods to infer the population structure of a set of individuals, by performing a cluster analysis to detect clusters (populations) and estimate the proportion of ancestry from each cluster (population) imagining that the genome of a given individual can be divided into segments of definite ancestral origin (J. and Lange 2009). There are two approaches to estimate global ancestry: model-based and non-parametric estimation. Non-parametric approaches use cluster analysis to seek clusters in the data without any population genetics model, using methods such as principle component analysis (PCA) and multidimensional scaling (MDS) to explain the variation between individuals at variable sites (J. and Lange 2009). EIGENSTRAT (Price et al. 2006; Patterson, Price, and Reich 2006) is one of the softwares that implements a PCA approach. The model-based approach is the most used in population genomics of both modern and ancient populations, especially the method implemented in *ADMIXTURE* (J. and Lange 2009) software e.g. (Lazaridis et al. 2017; de Barros Damgaard et al. 2018; Allentoft et al. 2015; Hofmanová et al. 2016; Mittnik A 2014; Skoglund et al. 2012). Model-based approaches estimate ancestry coefficients as parameters of simple population genetics models. Besides *ADMIXTURE,* such models are implemented in for example *FRAPPE* (Tang et al. 2005) and *STRUCTURE* (Raj, Stephens, and Pritchard 2013). There are other methods like NGSAdmix (Skotte, Korneliussen, and Albrechtsen 2013) that calculate the admixture estimates based on genotype likelihoods, accounting for low coverage data, but it is required to have the raw data for all samples we analyze. *STRUCTURE* uses a Markov Chain Monte Carlo (MCMC) to sample the posterior distribution of the allele frequencies of each population, and the proportion from each cluster for each individual (also called ancestry proportions). *ADMIXTURE* implements the same model but uses a maximum likelihood approach. This allows to apply it to large SNP datasets because 'high-dimensional optimization is much faster than high-dimensional MCMC' (J. and Lange 2009). The statistical model underlying *ADMIXTURE* requires unrelated individuals (*I*) genotyped at different SNPs (*J*), which are assumed to be neutral, and hence to reflect the demographic history of populations. The genome of each individual is considered to be drawn from *K* ancestral populations. Each ancestral population (*k*) contributes a fraction to an individual genome (q_{ik}) and each allele at each SNP has a frequency (p_{kj}) in population *k* (J. and Lange 2009). *ADMIXTURE* records the genotype data as counts, with g_{ij} representing the number of copies of one allele at a marker *j* of an individual *i*. It then estimates the parameters matrices $Q = \{q_{ik}\}$ and $P = \{p_{kj}\}$. The total number of parameters estimated is (I x K) + (K x J). If we have, for example, 2,000 individuals, 100,000 SNPs and assume 5 ancestral populations the total number of parameters to estimate is 510,000, which makes using optimization methods that involve manipulation of matrices computationally infeasible for high values of *K* (J. and Lange 2009). For this reason, *ADMIXTURE* uses two optimization methods: EM algorithm (Dempster et al. 1977), implemented in FRAPPE, to quickly reach the maximum vicinity and then changes to a faster block relaxation algorithm, because EM algorithm is very slow to converge (J. and Lange 2009). *ADMIXTURE* also allows to compute standard errors using a moving block bootstrap approach. Instead of resampling individual SNPs, it resamples blocks containing *h* consecutive SNPs (default *h* = 10cM genetic distance) with replacement and uses prior Q and P as starting parameters. It then computes standard error estimates instead of confidence intervals (J. and Lange 2009). The choice of *K* affects the runtime, with it increasing considerably with higher values of *K*. *STRUCTURE* uses a different bootstrap method by implementing a MCMC approach and outputs confidence intervals instead

of standard errors. *ADMIXTURE* bootstrap method has a speed advantage over the one implemented in *STRUCTURE* (J. and Lange 2009). For all the reasons stated above, and because it is the most used software in ancient DNA studies, we used *ADMIXTURE* to compute individual based estimates. *ADMIXTURE* allows to compute a Cross-validation error for each run, by dividing all observed genotypes into equal size folds (five by default). For each fold in turn it converts all the genotypes in that fold into missing data (masked dataset) (Alexander and Novembre 2015). It then uses other folds as a training set to fit a model which we then evaluate at the masked dataset (Parang, Wiebe, and Knaus 2012). This way we estimate the parameters $Q^{\wedge}$ and $P^{\wedge}$ for each fold. The genotypes in each fold are then predicted by the expected value:

$$E\left[g_{ij}\middle|\hat{Q},\hat{P}\right] \quad (3.1)$$

The mean squared error is calculated for each fold is then averaged across all folds to obtain the CV-error (Parang, Wiebe, and Knaus 2012).

3.2.2 Workflow

The dataset used for the population structure analysis and to infer the admixture estimates was adapted from (Lazaridis et al. 2017) (https://reich.hms.harvard.edu/sites/reich.hms.harvard.edu/files/inline-files/MinMyc.tar.gz), as is described both in this and the next section, 3.2.3. We obtained the BAM files (see section 3.1) from the six Bronze Age Greek individuals from our collaborators at the University of Lausanne. To compare this WGS data with other modern and ancient individuals we used a large SNP array panel of modern and ancient samples, obtained using the Human Origins Affymetrix array (https://reich.hms.harvard.edu/sites/reich.hms.harvard.edu/files/inline-files/Data.tar) (Patterson et al. 2012). This array has been developed to study human demographic history, aiming to include only neutral SNPs rather than SNPs in genes, so that demography is the main factor affecting variability. Due to the large number of individuals we ended up using (2,399), the use of this SNPs array is better than merging our WGS data with other raw sequencing data from WGS studies. This array has 1,237,207 SNPs for analyzing only Ancient genomes and 594,424 SNPs for Modern and Ancient joint analysis (info about the panel available at: https://reich.hms.harvard.edu/sites/reich.hms.harvard.edu/files/inline-files/Data.tar),. Our collaborators in University of Lausanne downloaded a Variant Call Format (VCF) file which had 2068 modern samples (https://reich.hms.harvard.edu/sites/reich.hms.harvard.edu/files/inline-files/NearEastPublic.tar.gz) genotyped on the Human Origins Array (594,424 SNPs) and downloaded the 351 ancient individuals from (Lazaridis et al. 2017) genotyped on the 1,237,207 SNPs of the Human Origins (from: https://reich.hms.harvard.edu/sites/reich.hms.harvard.edu/files/inline-files/MinMyc.tar.gz) and merged them to obtain a total of 621,272 SNP . The SNP array data from the panel is stored in VCF format, but no available tool exists to merge WGS BAM files and VCF files. Before SNP filtering, we used VCFTools v.0.1.17 (Danecek et al. 2011) to remove all ancient samples and only keep modern individuals, as ancient individuals can have more damage than modern individuals and hence have more sequencing errors. For each software we show a table with the command line options used within the software and each option description. Below we have the input used to filter modern samples from our dataset (Table 3.2), with VCFTools (Danecek et al. 2011):

Table 3.2 VCFTools command and flag description for removing individuals from a VCF file

vcftools –vcf input –remove remove.list –recode –out output

--vcf	*Input VCF file*
--remove	*Input List of ancient samples*
--recode	*Flag to recode our VCF*
--out	*Name of the VCF to output*

Then, we used Plink v.1.9 (Chang et al. 2015) to filter SNPs with 0.05 or more minor allele frequency (Table 3.3) to exclude rare variants:

Table 3.3 PLINK command and flag description to filter by maf of 0.05

plink –vcf input –maf 0.05 –recode –out output

--vcf	*Input VCF file*
--maf 0.05	*Minor Allele Frequency*
--recode	*Flag to recode our VCF*
--out	*Name of the VCF to output*

From the original number of SNPs, after this step 428,046 SNPs were kept. To keep only independent SNPs, we then performed Linkage Disequilibrium (LD) pruning using the same parameters as (Lazaridis et al. 2017): a 200Kb sliding window shifting every 25 variant counts (Table 3.4). All SNPs with r^2 bigger than 0.4 were removed which led to 165,447 final SNPs (higher r^2 between two sites suggest higher observed frequency of two alleles in those sites comparing to the expected frequency if the two alleles segregated independently, suggesting high linkage between those sites). It outputted two lists, one with sites to retain and the other sites to exclude:

Table 3.4 PLINK command and flag description to remove sites in linkage

plink –vcf input –indep-pairwise 200 25 0.4

--vcf	*Input VCF file*
--indep-pairwise 200 25 0.4	*LD Parameters*

With the newly generated list of independent SNPs to retain, we made a BED file to later use in VCFTools (Danecek et al. 2011) to filter the original dataset (ancient + modern) for those sites (Table 3.5). The BED file is tab-separated with three columns: Chromosome number, the position of our SNP minus 1 and the position of our SNP. We used the generated BED file to filter the dataset for those positions:

Table 3.5 VCFTools command and flag description extracting positions from a VCF file using a BED file

vcftools –vcf input –bed keep.bed –recode –out output

--vcf	*Input VCF file*
--bed	*BED file of sites to keep*
--recode	*Flag to recode our VCF*
--out	*Name of the VCF to output*

From the filtered VCF, we removed nineteen Greek Samples from Lazaridis, with VCFTools (Table 3.2) (Danecek et al. 2011), so that all Bronze Age Greek samples on the study (both Lazaridis's and ours) go through the same procedures and are added as a whole to the VCF file, as detailed in 3.2.3.

All low coverage ancient samples in the reference panel are coded as homozygous. This is typically the case for ancient DNA datasets with low coverage. When the coverage is low, rather than calling genotypes, the approach used is to randomly sample a single read at each position of the genome for each individual. Thus, at each position, individuals can only have one copy of the reference or alternative allele, which are coded as homozygous reference or homozygous alternative. No software or tool is publicly available to draw one allele at random. For that reason, we developed a script in R to randomly draw one read for each site for each one of our BAM files. We also did a script to remove transitions from the VCF file, which resulted in a final dataset with 30,896 SNPs. Because it was a low number of sites, we chose to use both transitions and transversions. Besides the six samples and the nineteen ones from Lazaridis, we added four extra high-quality ancient genomes to the panel because they were used in other analysis performed by my colleagues. So, a total of twenty-nine BAM files had to be merged with the dataset.

3.2.3 Sampling one read from BAM files and merging with a VCF

As described above, for ancient DNA with low depth of coverage, rather than calling genotypes, the commonly used approach is to randomly sample a single read at each position of the genome for each individual because of the low coverage associated with ancient genomes. However, there is no standard program to perform this. Thus, we developed a bash script that uses several programs to implement this sampling method which was applied on the BAM files (remapped for the human genome build 37.1) that we wanted to add to our analysis (Table 4.1). Those BAM files were our 6 WGS samples, 19 ancient SNP array samples from (Lazaridis et al. 2017) , two modern WGS from (Mallick et al. 2016), two ancient WGS from (de Barros Damgaard et al. 2018), one WGS genome from (Jones et al. 2015) and one WGS genome from (Hofmanová et al. 2016)

We started by using ANGSD (Korneliussen, Albrechtsen, and Nielsen 2014) to get an allele count matrix where the number of rows is equal to the number of SNPs and each individual is represented by four columns, where each column is the number of times that a given allele appears on the reads. Before the allele sampling, the reference and alternative allele for each position in the VCF were extracted (RefAlt file) using awk:

awk '{print $1, $2, $4, $5}' dataset.vcf > dataset.refalt *(3.2)*

Then, we made an R script to sample one allele for each individual from the previous matrix according to the allele frequencies in each position. By comparing the drawn allele with the ones in the RefAlt file we coded the extracted allele as 0/0 or 1/1 if they matched the reference or alternative, respectively, or ./. if it did not. The output is a matrix with *N* columns, where *N* is the number of individuals and *S* rows, where *S* is the number of SNPs. Then with shell scripting the produced matrix was merged with the reference panel. Some modern genomes, with high coverage, in the panel had heterozygous sites, meaning 50% of the reads had the alternative and 50% the reference allele. For that reason, we sampled one of the alleles at random at each site (Figure 3.3 has a schematic representation of the method),

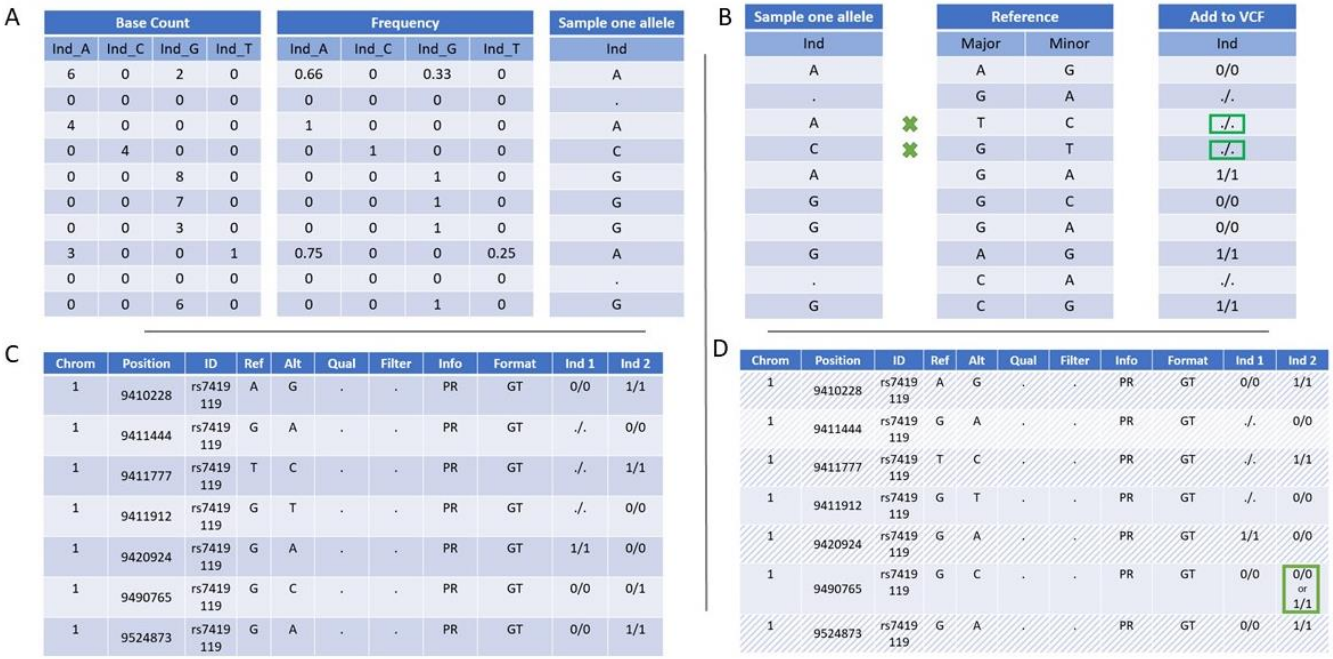

Figure 3.3 Scheme of our method for sampling one read. A) From ANGSD output get the base count matrix and compute the frequency matrix to than sample one allele. B) Compare extracted allele with the reference to code it properly. C) Merge with the VCF file and D) Extract one of the alleles for any positions that may be heterozygous in the panel

With VCFTools (Danecek et al. 2011) we calculated the missing data per individual (Table 3.6) and twenty individuals (three from (Lazaridis et al. 2017) BAM's) that had over 95% of missing data were removed (Table 3.2).

Table 3.6 VCFTools command and flag description to output the amount of missing data in all individuals.

vcftools –vcf input–missing-indv	
--vcf	*Input VCF file*
--missing-indv	*Output amount of missing per individual*

The final dataset for the admixture analysis had 2,399 Individuals (2,068 Modern and 331 Ancient) and 165,447 SNPs. Because ADMIXTURE (J. and Lange 2009) uses Plink files (ped + map) I converted the VCF into the required formats with Plink (Chang et al. 2015) (Table 3.7):

Table 3.7 PLINK command and flag description to recode a VCF to Ped/Map formats

plink –vcf input –recode12	
--vcf	*Input VCF file*
--recode12	*Flag to recode as Ped file*

We used the software *ADMIXTURE* (J. and Lange 2009) to estimate the ancestry proportions for each cluster for each individual. The software takes a set of SNPs from unrelated individuals and models the

probability of the observed genotypes using ancestry proportions and population allele frequencies from *K* original populations. It maximizes the likelihood of the fraction that each *K* population has contributed to the genome of each individual (J. and Lange 2009). It outputs three files with the estimates, a *Q* matrix with the inferred ancestry proportions for each individual, a *P* matrix with the estimated allele frequencies at each SNP for each cluster and a log file from where we can obtain the Cross-Validation error (CV-error). For each value of *K* ranging between 2 and 17, we performed 10 run replicates using the cross-validation flag (Figure 4.2). Because we used data from sampling one random read at each site for each individual, we ran *ADMIXTURE* with the haploid flag (*--haploid="*"*). Then, we used a R script to order the columns of the *Q* files so that the major components of each population could be assigned to the same color to obtain a clearer figure. Then, to answer our objectives of the thesis, we plotted all Eurasian ancient samples and modern Greeks, Cretans and Cypriots, organized by age for all the *K* values [Supplementary Figure 1]. The chosen run for each *K* was the one with lowest CV-error (Figure 4.2). We used the R package *popHelper* (Francis 2017) to better illustrate the results and make the visualization simpler. Based on the low CV-error and the fact that Neolithic Farmers and some Hunter-Gatherers populations formed their own clusters of almost 100% ancestry, we chose *K = 11* to make a plot with a subset of individuals belonging to each relevant population (Figure 4.4). The amount of missing data in all our samples varies considerably (Table 4.1) and to verify how missing data correlates with the error estimation of each component we did a bootstrap of 200 replicates performed for *K=11* (using the flag -b in *ADMIXTURE*). Then we plotted the standard error bars associated with the four main components for each one of the samples (Figure 4.5).

3.3 Inbreeding

Mating between closely related individuals is common among humans, at least 10% of the world population have parents that are at least second-degree cousins, which may be due to cultural practices or due to a small population size (Ceballos et al. 2018). Individuals from isolated populations, by having a lower effective size, tend to be more closely related to each other and may inherit identical chromosome segments from their parents. This results in uninterrupted long runs of homozygous genotypes – Runs of Homozygosity (ROH) (Ceballos et al. 2018). A higher number of longer ROHs are often associated with increased risk of schizophrenia, Alzheimer, autism, some types of cancer and coronary heart disease, among other health problems (Ceballos, Hazelhurst, and Ramsay 2018). Populations in islands are likely more isolated than ones in mainland and hence might have low effective sizes. A way of measuring how isolated and inbred one population is can be done by estimating the number and the length of ROHs. For example, the proportion of endogamous people from an isolated Orkney island that have ROHs over 10Mb (~ 30%) is higher than those from mainland Scotland (~1%) (McQuillan et al. 2008). In order to detect Runs of Homozygosity in the six Bronze Age individuals from Greece, we restricted our analysis to chromosome 21. To estimate homozygosity only at positions that are variable, we filtered our BAM files for the Human Origins SNP array for chromosome 21. We did not analyze all autosomal chromosomes due to time constrains. Thus, we choose to use chromosome 21 since it is the shortest. We filtered the BAM files to extract the reads that encompass our SNP, using SAMtools v.1.9 (Li et al. 2009) (Table 3.8).

Table 3.8 SAMTools commands and flag description to filter BAM files reads that overlap the positions in a BED file

samtools view -b -h -L variants.bed individual.bam > individual_variants.bam

-b	*Output in Bam format*
-h	*Include the header*
-L	*Output alignments overlapping the positions in the Bed file for SNV*

Then we used ANGSD (Korneliussen, Albrechtsen, and Nielsen 2014) to call genotypes with the standard parameters and code it in Plink files, converting the BAM files into tped and tfam file formats. With the program Plink (Chang et al. 2015) we used input options that are proven to be good to find ROHs in low coverage data (Ceballos, Hazelhurst, and Ramsay 2018): a ROH had to have a minimum of 50 SNPs; each sliding-window had 300kb; a minimum of 1 SNP at each 50kb was required to be considered a ROH; a maximum length of 1000kb between SNPs in order for them to be considered at two different segments; between three to five heterozygous SNPs were allowed to be in each sliding-window. There are other software's that calculate ROHs using Hidden Markov Models, like BCFTools (Narasimhan et al. 2016), but require either a file with allele frequencies at each SNP at the population from where the individual was sampled or a VCF with high number of individuals from the population our individuals belong to, to call ROHs. Considering that almost no ancient genomes are whole genome sequenced and that we are the first ones to do it in Early Bronze Age Greeks, these methods are hard to implement. For this reason, we developed a new method attempting to detect ROHs on samples with low depth of coverage, by using genotype likelihoods to find windows with a high probability of homozygosity across all sites. Using the filtered BAM files for chromosome 21 we used ANGSD (Korneliussen, Albrechtsen, and Nielsen 2014) to

generate a beagle file with the genotype uncertainty (Table 3.9). The Beagle file is a matrix with a column of the position in the chromosome, a column with the reference and alternative allele and, for each individual, we have three columns, the first one has the likelihood of being homozygous for the reference allele, the second has the likelihood for being heterozygous and the third one the likelihood of being homozygous for the alternative allele. The genotype likelihoods were calculated with SAMTools (Li et al. 2009) model implemented in ANGSD (J. and Lange 2009):

Table 3.9 ANGSD commands and flag description to calculate genotype likelihoods for all BAM files

angsd -GL 1 -out ind_genolike -bam ind.bamfilelist -doGlf 2 -doMajorMinor 1

-bam	*Bamfile list*
-GL 1	*Samtools model for Genotype Likelihood*
-doGlf 2	*Output in beagle likelihood file*
-doMajorMinor 1	*Infer Major and Minor allele from GL*
-out	*Output file name*

My script takes the BEAGLE file and gets a matrix with the chromosome number, position and the probability of being homozygous, then for a given chromosome it checks the sites where the probability of being homozygous (*P*) is equal or higher than a given threshold (ex: $P >= 0.8$). Consecutive sites with *P* above the threshold were considered to be in a potential ROH. We also allowed sites with, for example, $P > 0.5$ to be in a ROH if they were between sites with $P >= 0.8$ (Scheme of the method in Figure 3.4). The output of my script is a list of ROH in Megabases (Mb), over 0.1 Mb, for a chromosome of each individual. The length of each ROH is calculated given by the following expression:

$$ROH = \left(\frac{Pos.\,Last\,T - Pos.\,First\,T}{1{,}000{,}000}\right) Mb \tag{3.2}$$

$$eg\text{: } ROH = \left(\frac{9{,}420{,}924 - 9{,}410{,}228}{1{,}000{,}000}\right) \sim 0.01 \text{ Mb} \tag{3.3}$$

A

Chr_Pos	Allele 0	Allele 2	Ind_HH	Ind_Hh	Ind_hh
21_9410228	2	0	0.969698	0.030302	0.000000
21_9411444	1	0	0.969698	0.030302	0.000000
21_9411777	3	0	0.984616	0.015384	0.000000
21_9411912	3	0	0.666580	0.333287	0.000133
21_9420924	0	1	0.984616	0.015384	0.000000
21_9490765	2	0	0.666580	0.333287	0.000133
21_9524873	0	1	0.000797	0.998406	0.000797

B

Chr	Pos	1 - Ind_Hh		0.5 > P > 0.8	P > 0.8	Union	
21	9410228	0.969698	✚	F	T	T	ROH
21	9411444	0.969698	✚	F	T	T	
21	9411777	0.984616	✚	F	(T)	T	
21	9411912	0.666713	●	T	F	T	
21	9420924	0.984616	✚	F	(T)	T	
21	9490765	0.333287	✖	F	F	F	
21	9524873	0.001594	✖	F	F	F	

Figure 3.4 Scheme of our ROH calculation method. From a matrix with the genotype likelihoods (A), we check whether or not we have successive sites that meet our criteria of calling ROH (B) and generate a matrix with the union of the sites that belong to a ROH in order to compute the length.

4. Results

4.1 Data

The resulting dataset of merging the WGS with SNP array data comprised 621,272 SNPs and 2399 individuals. Regarding the missing data per individual at those sites, as expected, ancient individuals have a much higher proportion of missingness than modern samples. Missing data in modern samples was less than 6.5% (mean 0.5%), whereas for ancient individuals, the average proportion of missing data was 48.01% (Figure 4.1).

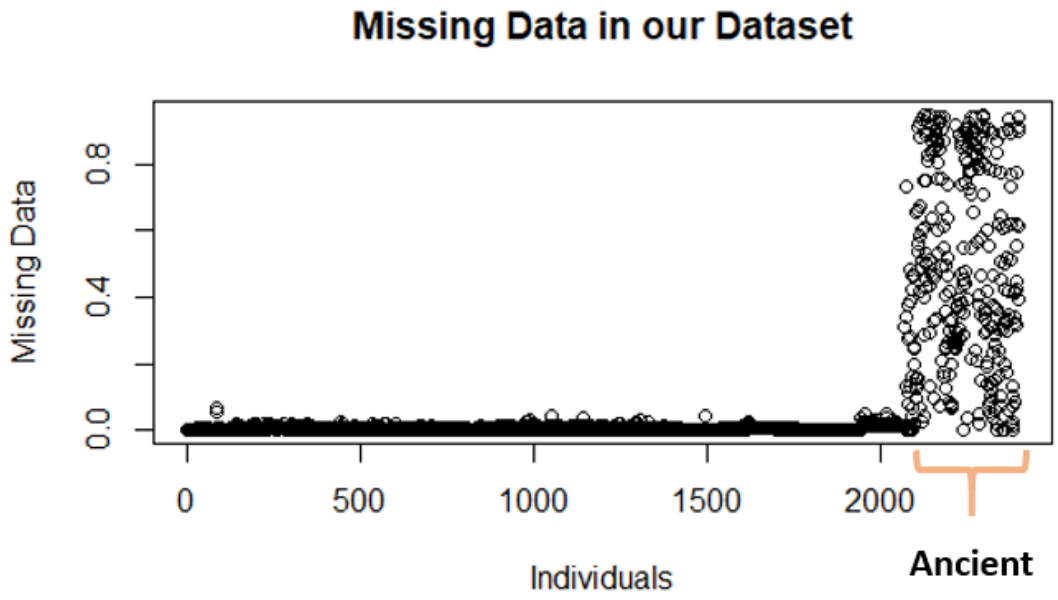

Figure 4.1 Missing data measured as the proportion of SNPs without data for each individual. Ancient samples have a much higher proportion of missing data than in modern samples. This is expected as ancient DNA is less preserved and is more degraded.

Compared with other ancient samples, our six Bronze Age individuals have a lower amount of missing data, between 2.0 and 13.0%, whereas (Lazaridis et al. 2017) BAM files have missing data values ranging between 7.91 and 93.5% (Table 4.1). The extra four genomes included are high quality ones, two of them are modern from Crete and Greece with 0.03% and 0.04% of missing data, respectively. The Yamnaya individual has 0.08% and the Easter Hunter-Gatherer 12.9 % of missing data. In Table 4.1 we have the missing data for the BAM files that were added to the original dataset and kept for the analysis.

Table 4.1 Proportion and total Number of missing SNPs in our BAM files. Three BAM files had over 95% missing data and were not used for any analysis.

Individual	Population	Number of SNPs	Number Missing SNPs	% Missing SNPs
KK1.SG	Caucasus Hunter-Gatherer	165447	212	0.13
EHG_Sidelkino_deBarros18	Easter Hunter-Gatherer	165447	21406	12.94
Bar8.SG	Anatolia N	165447	1407	0.85
I2937.1240K	Greece N	165447	88627	53.57
YamnayaKaragash_EBA	Yamnaya (Steppe EMBA)	165447	127	0.08
I2495.1240K	Anatolia BA	165447	66344	40.10
I2499.1240K	Anatolia BA	165447	126145	76.24
I2683.1240K	Anatolia BA	165447	48762	29.47
Mik15	EBA Helladic	165447	16451	9.94
Kou01	EBA Cycladic	165447	21440	12.96
Kou03	EBA Cycladic	165447	12015	7.26
Pta08	EBA Minoan	165447	4089	2.47
I0070.1240K	Minoan Lasithi	165447	54075	32.68
I0071.1240K	Minoan Lasithi	165447	13092	7.91
I0073.1240K	Minoan Lasithi	165447	51296	31.00
I0074.1240K	Minoan Lasithi	165447	66288	40.07
I9129.1240K	Minoan Odigitria	165447	154838	93.59
I9005.1240K	Minoan Lasithi	165447	68882	41.63
I9130.1240K	Minoan Odigitria	165447	150043	90.69
I9131.1240K	Minoan Odigitria	165447	148674	89.86
Log02	MBA Helladic	165447	3331	2.01
Log04	MBA Helladic	165447	1914	1.16
I9006.1240K	Mycenaean	165447	73160	44.22
I9010.1240K	Mycenaean	165447	101092	61.10
I9033.1240K	Mycenaean	165447	99477	60.13
I9041.1240K	Mycenaean	165447	64392	38.92
SAMEA3302625	Greek	165447	68	0.04
SAMEA3302765	Crete	165447	60	0.04

4.2 Population Structure – Individual based admixture estimates

ADMIXTURE infers for each individual the proportion of the genome belonging to *K* clusters and the allele frequencies for each SNP in those clusters (J. and Lange 2009). We run *ADMIXTURE* for *K* ranging between two and seventeen for all the 2399 individuals (SI_2), and for each *K* we did ten runs and chose the one with lowest Cross-Validation error (CV-error) for each *K* (Figure 4.3). All the admixture plots shown in the manuscript are just a subset of the admixture plot in SI_2

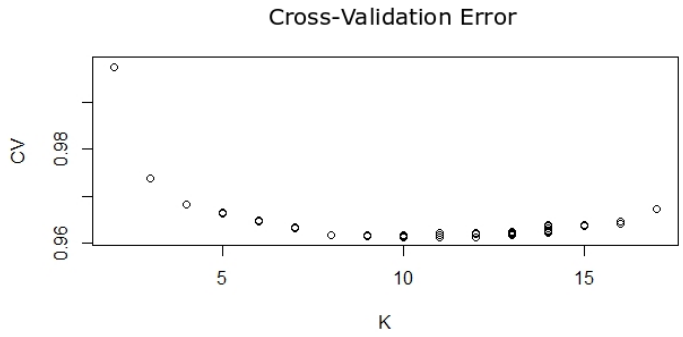

Figure 4.2 Cross-Validation error plot with lower CV-errors for K ranging from 8 and 13.

We selected ancient individuals from the Aegean and surrounding areas and modern Greeks, Cypriots and Cretans to be shown in the Admixture plot (Figure 4.4 and Supplementary Figure 1) because they are the ones that are relevant for this study. Usually, when dealing with modern data, *K = 2* shows two distinct clusters (usually Africans vs Non-Africans) because of the higher genetic variation among Africans compared to non-Africans. In our analysis the results for K=2 does not reflect this, as can be seen in Figure 4.3 A, which shows the ancestry proportion for a sub-set of individuals from Africa, Europe, Asia and Americas. Instead, for K=2 our estimates cluster Africans and most Eurasians together, and separates these from East Asian and American populations. Focusing on a sub-set with 25 of the 2,399 individuals, we see this in Figure 4.3, as Greeks (European) and Mota (African) have similar proportions, which are separated from Chukchi (Siberia) and Piapocco (Venezuela and Colombia, America). At K = 3 and at K = 4 we already recover the usual African vs Non-African clustering, and we have four groups that are well separated: Africans, Europeans, Americans and Asian. If we remove transitions (Figure 4.3 B), we obtain the same results. To check the effect of the MAF filter, we run *ADMIXTURE* on the same dataset with 2,399 individuals without applying the MAF filter (Figure 4.3 C). In that case, we see that for K=2, Africans are clustered together. Clearly, the effect of the MAF filter might be due to the overrepresentation of non-Africans in our panel.

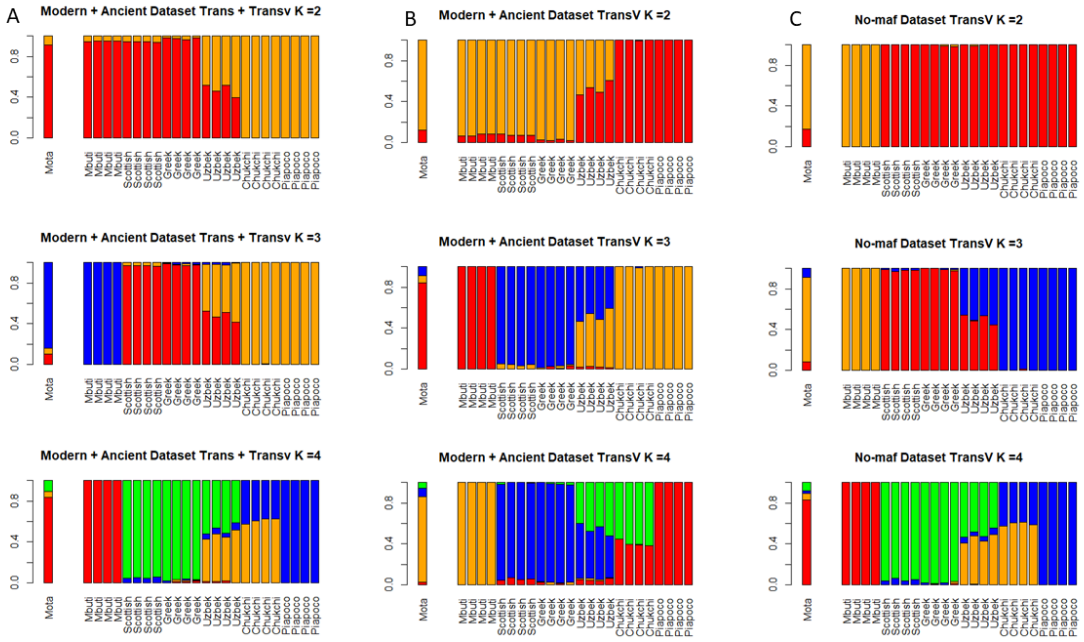

Figure 4.3 Admixture plot for three different datasets which shows the ancestry proportion for a sub-set of individuals from Africa, Europe, Asia and Americas: Two datasets where we applied the MAF filter: one with Transitions and Transversions – 165,447 SNPs(A) and the other only with Transversions – 30,896 (B); the other dataset with no maf filtering and only transversions –55,717 SNPs (C).

Starting from *K = 8* (Supplementary Figure 1) we already see some genetic structure within the sub-set of 374 individuals from the area we focused on: populations from the Caucasus region are majorly formed by a blue component; Hunter-Gatherers from Europe are formed by an orange component; and Neolithic Farmers from Greece and Anatolia show a red component. Although our six Bronze Age samples from Greece are estimated to have a major component related to Neolithic Farmers, they also show two more

components that are shared with the other individuals from the Chalcolithic (end of Neolithic) onwards (except for European MNChl and Levant). For *K* = 11 (Figure 4.4) we see a new cluster grouping together Natufians (-67%). This, together with the above mentioned three clusters, are the four most relevant components to explain our data. We see that populations from a given geographical location start to show components from other regions as we go through time along the Bronze Age, suggesting that Bronze Age was a time where Aegean people moved and admixed with each other. For instance, EN and EMBA Steppe populations have two major components: European HG and Caucasus/Iran component, but in MLBA they start to show Neolithic Farmer component. This is seen also when looking at Anatolians, as there is an increase of Caucasus/Iran components when comparing Chalcolithic and BA samples to older Neolithic samples. Our Logkas individuals differ from the Early Bronze Age samples as they show, for *K* > 5 (Supplementary Figure 1), a higher ancestry proportion of the European Hunter-Gather component. Interestingly, Modern Greeks are very similar to our Lokgas samples, but with higher proportion of Caucasus and Natufian components, indicating that the beginning of the Bronze Age started to close the genetic gap between modern and ancient Greek individuals. We chose *K=11* as the most robust to visualize the results, as it is one of K values with lower Cross Validation (Figure 4.2) and because for *K*=11, there are four well defined clusters that correspond to Neolithic Farmers, European HG, Iran/Caucasus-related populations and Natufians-related. The plot for *K* = 11 (Figure 4.5) and *K* = 2-17 (Supplementary Figure 1) is organized by age. There are four main components: Neolithic Farmer (red) present in Neolithic Anatolians and Greeks (~95%); Iranian/Caucasus (blue) related component in Iranian Neolithic (~89%) and Caucasus Hunter-Gatherers (~70%); European Hunter-Gatherer (orange at approximately 100% in Western Hunter-Gatherers and Scandinavian Hunter-Gatherers), which is also a Steppe-related component; and Natufian component (brown ~69%). Bronze Age Anatolians and Greeks (Mycenaeans and Minoans) show an extra component that was absent from Neolithic samples. They have a Caucasus component and a residual amount of Natufian component. Our Early Bronze Age individuals (Kou01, Kou03, Mik15 and Pta08) are more similar to Minoans than to Mycenaeans. This was expected for the island samples (Kou01, Kou03 and Pta08), especially for Pta08 as Minoans are also from islands (the same one as Pta08 - Crete). Mycenaeans have one extra component that was not estimated for any of our EBA samples: the European HG component. The admixture estimates, as can be seen in Figure 4.4 indicates that Mik15 has the European HG component. The bootstrap method applied afterwards (Figure 4.5) confirms that Mik15 has the European HG component (as the error bar does not reach zero), despite it being very low. As we move forward in time, the first samples where we detect individuals with high proportion of this European HG component in Greece, are Log02 and Log04. Log02 has an estimate of 22% of European HG, while Log04 has 28%. These samples are also very similar to modern Greeks and Crete individuals, only lacking the Natufian-like component in similar proportions. We also see that Steppe populations start to have the Neolithic Farmer component in Middle and Late Bronze Age. Modern Cypriots differentiate from the other modern individuals and our Log02 and Log04 by having an extra Natufian and Caucasus components and lacking the European HG component (except for one of them). Comparing ancient individuals with high amount of missing data (ex: Minoan Odigitria) and others with low amount (ex: our six individuals) we see that there is little difference in the error interval bars (Figure 4.5),

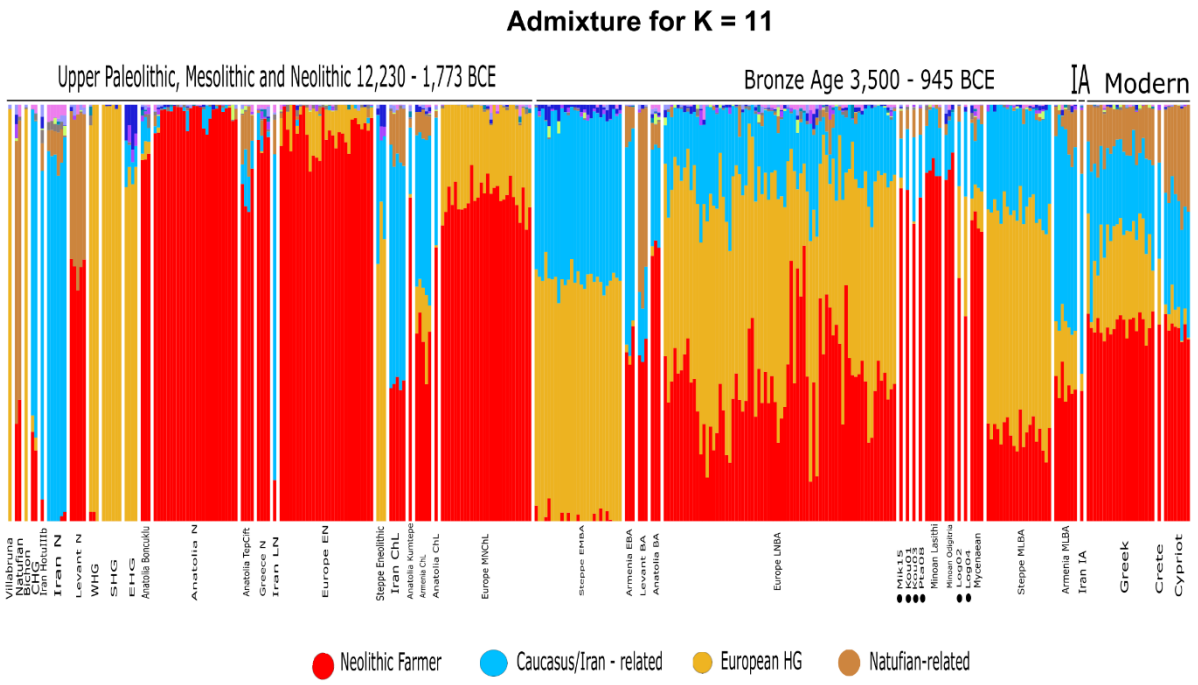

Figure 4.4 Admixture Plot for K = 11 for a subset of 374 Individuals (from the merged SNP array and WGS 2399 total samples) of each relevant population. This was done with 165,447 SNPs from the Human Origins SNP Array. The red component suggests a Neolithic Farmer ancestry; blue – Caucasus/Iran-related ancestry; Orange – European HG (steppe-related) ancestry; brown – Natufian related ancestry. Populations are ordered according to age and our samples have a black dot behind their label

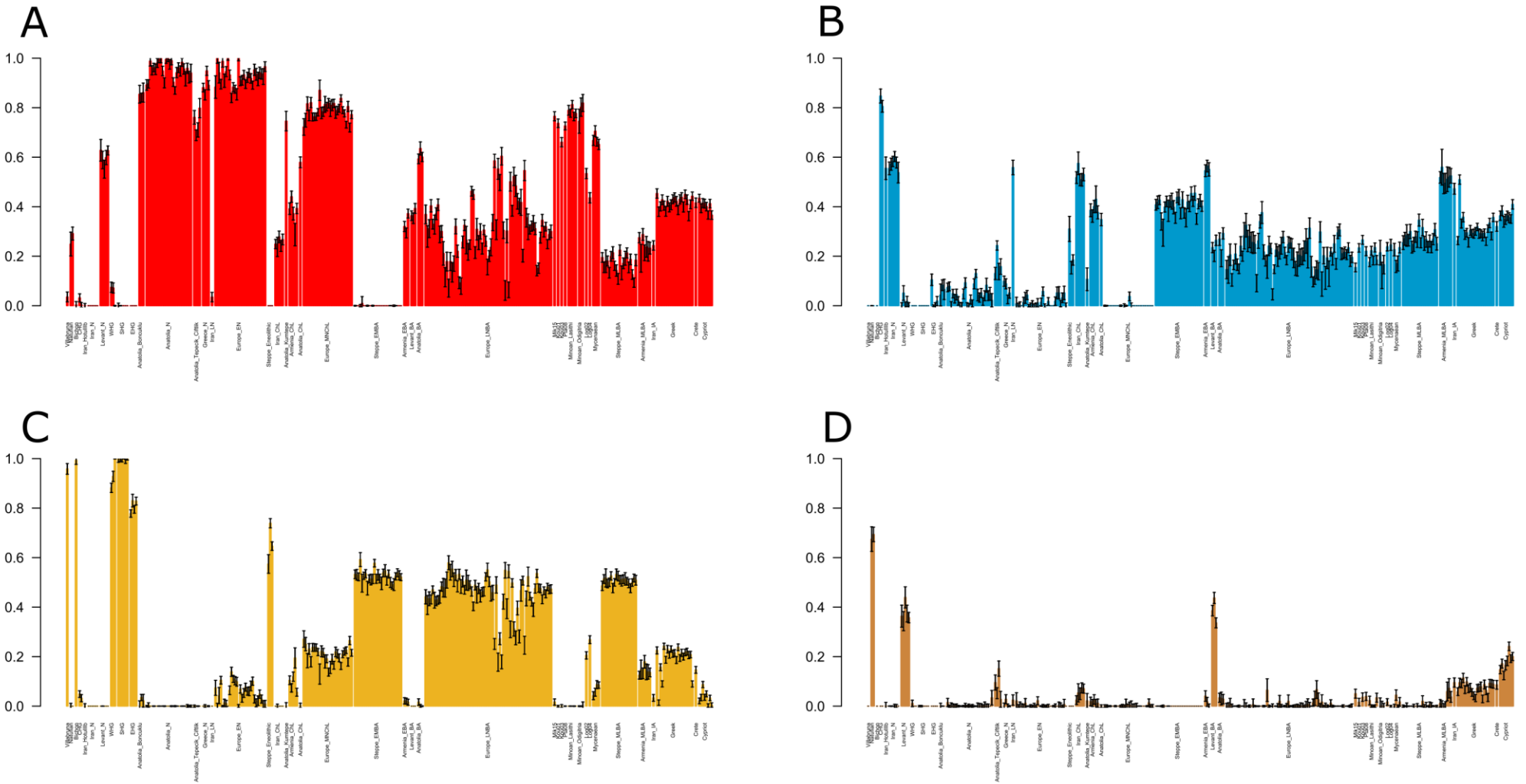

Figure 4.5 Bootstrap error bars for the same data of Figure 4.4 for the four main components, in the following order: A) Neolithic Farmers, B) Iran/Caucasus-related, C) European HG and D) Natufian-related.

4.3 Inbreeding

When we applied the PLINK method for estimating Runs of Homozygosity, we did not manage to find any ROH on Chr21 (we used 300 Kb to be the threshold as (Ceballos, Hazelhurst, and Ramsay 2018) used for low-coverage data). Using our newly developed method, we managed to detect only small ROHs for Chr21 in our samples (applied same thresholds that we used to exemplify our method in the Methodology section). We could not detect any ROH larger than 0.31 Mb on our Log02 chromosome 21. Even though no ROH of considerable size was found, our method works and detects stretches of homozygous genotypes. Even with the small ROHs we can compute the total number of ROHs (NROH) and the total sum of ROHs length (SROH), in Mb, for each individual. We observe that Log02 and Kou01 have more NROH over 0.1 Mb and a higher SROH then the other samples. We do not have any distinction between ROH patterns between individuals from the mainland and islands.

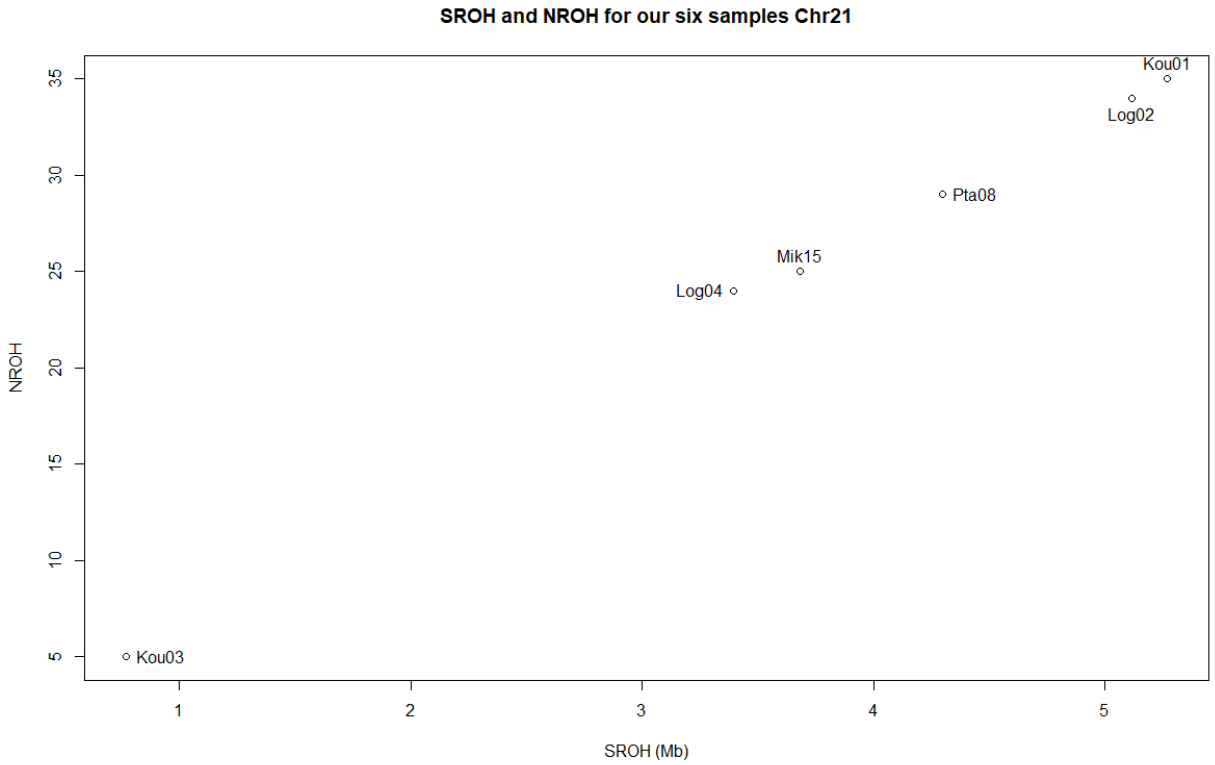

Figure 4.6 Plot with total sum length sum of ROH (SROH) and total number of ROH (NROH) of our 6 WGS samples, for the chromosome 21 SNPs in the Human Origins .

5. Discussion

In this study we analyzed genomic data from Neolithic and Bronze Age individuals from Eurasia, especially in the regions around Greece. We used newly sequenced Bronze Age ancient samples from mainland Greece (Log02, Log04 and Mik15), Crete (Pta08) and Koufonisis (Kou01 and Kou03) and compared them with other ancient and modern samples. We characterized the population structure based on model-based clustering and inferred the inbreeding levels based on the runs of homozygosity (ROH). To analyze the resulting low depth of coverage genomic data we developed two pipelines: (1) merge WGS with SNP array data, sampling one read at random from each SNP; (2) estimate ROHs for low coverage WGS data using a simple heuristic approach.

5.1. Merging samples - a common challenge with ancient DNA

To compare the Bronze Age samples from Greece with other ancient and modern samples, we merged them with the dataset from (Lazaridis et al. 2017). This seems a trivial procedure, but since Lazaridis et al (2017) used a SNP array and we used whole genome sequencing (WGS) data, and had a very low depth of coverage, we had to be very careful in merging our dataset with the existing panel from Lazaridis et al. (2017). Due to their relevance for our analyses, we also included four extra genomes: one modern Greek and one modern Crete (Telenti et al. 2016) and one ancient Yamnaya Karagash and one ancient Sidelkino Eastern Hunter-Gatherer (de Barros Damgaard et al. 2018). Datasets with ancient and modern DNA genotypes are stored in VCF formats, but if we sequence new samples there is no straightforward way of merging their BAM files with the publicly available VCFs, because BAM files store the aligned reads, whereas VCF only stores genotypes (usually only at variant sites). One solution to deal with ancient DNA when coverage is low and varies between individuals is to sample one read per SNP per individual. In those cases, ancient DNA data is coded as homozygous for either one of the alleles. In this case all individuals are comparable even if the depth of coverage is very low and differs across individuals. Thus, rather than calling genotypes, which will be associated with a high error for low coverage, one allele is sampled at random for each SNP. Although this is a common practice, we found no available tool or script to perform this operation. For that reason, we developed a script that takes a matrix with the number of reads that have a given allele for a given list of sites. We then sample one of the alleles at random, for each individual at every site, according to their frequency. By comparing the resulting allele with the reference and alternative allele for that position in the dataset, we decide to code that individual as homozygote reference or homozygote alternative allele. It then outputs a file which can then be added to the dataset we want to merge our data with (using shell scripting). By making this script publicly available in github we are the first ones, as far as we know, to provide a way of sampling one read from a BAM file and merging it with a VCF file. Here we merged the WGS and SNP array from a panel with many individuals across Eurasia and at sites that are neutral, since we aimed to infer aspects of population structure related with the demographic history of populations, minimizing the influence of natural selection on these patterns.

5.2. Data quality after merging WGS with SNParray data

When looking at the merged dataset we see a noticeable difference in missing data between ancient and modern samples (Figure 4.1). The degradation of ancient DNA and deamination makes it harder to have good quality genomes with a high depth of coverage, for instance from the (Lazaridis et al. 2017) SNP array genomes, that we added to the dataset, only two had less than 30% of missing SNPs (Table 4.1). Because of this high amount of missingness in ancient samples, we decided to: i) apply the MAF filter; ii) take linkage-disequilibrium (LD) into account to filter out linked sites to obtain a final data with independent SNPs, using the same parameters as (Mittnik A 2014; Lazaridis et al. 2017), on the modern individuals. The MAF filter is applied to remove rare variants that are just found in a few individuals, which are more likely due to sequencing errors. To apply this filter, we calculate the frequency of the alleles for each SNP, removing sites where we have rare alleles with a frequency lower than 0.05. MAF filtering removed 193,226 sites. If we happen to have oversampled populations, filtering by MAF could eliminate variable positions, not because they are sequencing errors but because they are specific to a particular population with few individuals represented in the dataset. That is what happened on our filtered dataset. Since we only have 123 African individuals in a sample size of 2399 individuals (~5 % Africans), when we applied a MAF of 0.05 we likely removed SNPs that are variable in Africa. This introduces a type of ascertainment bias, which can explain why our admixture results for K=2 does not separate Africans from non-Africans (Figure 4.3). However, we repeated the analysis without the MAF filter and found similar results for $K > 3$. This differences between datasets is only meaningful for $K = 2$, as we see that for $K = 3$ we have Africans, Europeans and Asians/Americans in both analysis with and without MAF filters, without major differences in ancestry proportions. We get the same results for $K = 4$ in all scenarios, but now with a separation between Asians and Americans. In sum, applying this MAF filter only has repercussions, within our dataset for $K = 2$ and ancestry proportions for K values above two have the same estimates. Since we are mostly focusing on Eurasian samples, this ascertainment is likely not important, and we preferred to still apply the MAF filter to remove rare variants due to sequencing errors.

5.3. Population Structure changes in Neolithic and Bronze Age

In this study we aimed to answer two main questions: How do the BA samples relate to other ancient and modern samples? How do the BA Greek samples relate with each other, are there differences due to culture? We performed the model-based clustering analysis implemented in *ADMIXTURE* assuming a different number of clusters, K. Here, we discuss the results for K=11 (on a subset of 374 individuals relevant to answer our questions) because at this K value we start to separate individuals into four clearly distinct clusters: the Neolithic Farmers (~95% in Anatolia_N and ~80% Greece_N), European Hunter-Gatherers (~ 100%% in WHG and SHG), Caucasus Hunter-Gatherers/Iranian Neolithic - related (~89% in Iran_N and ~70% in CHG) and Natufian -related (~67% in Natufians). The cross-validation plot also shows that eleven is one of the values for which the CV-error is lower, indicating it is one of the K values that better explains the data (Figure 4.1). It is noteworthy that the Steppe populations have a large contribution from the European Hunter-Gatherer cluster, and for that reason we call this the “steppe-related” component.

Interestingly, there is variation in the proportion of these 4 clusters across individuals, which could be interpreted in terms of spatial population structure and temporal changes in population structure. Starting with the oldest samples from the Neolithic, we found that the Levant individuals, one of the first population

of farmers, have a high proportion of the Natufian-related cluster (~33%) (Natufians are a 11,840-9,760 BCE individuals from a semi-sedentary population, with some type of agriculture, from the region of present-day Israel). They also have the Neolithic Farmers component. This component is maximized in Neolithic Anatolian 6,500-6200 BCE from present-day Turkey and that already practiced agriculture. These results are in accordance with (Lazaridis et al. 2016), which point out that Levant people have ancestry from Natufians and European Farmers. European Hunter-Gatherer groups form another cluster, which is shared with Steppe populations. In the Steppe samples from early Neolithic (Steppe_EN), in comparison with individuals from Europeans HG, we estimated an extra Caucasus-like ancestry correlated with Caucasus Hunter-Gatherers (CHG), which was also found by (de Barros Damgaard et al. 2018). When focusing in Greece, our admixture estimates indicate that Neolithic Greeks share the same major component with Neolithic Anatolians and Early Neolithic Europeans (from Spain, Germany and Hungary). The fact that Anatolians, Greeks and other Europeans share the same component in such high proportion is in accordance with (Hofmanová et al. 2016) that suggests a route through the Aegean as one source of agriculture introduction in Europe.

In the Bronze Age we found that Greek samples (Mycenaeans, Minoans and our six samples) are different from Neolithic Greeks, with genomes that could be modeled as a mixture of two or more clusters, with the predominant Neolithic Farmer and Iran/Caucasus components. All newly sequenced six Bronze Age (EBA) samples from Greece (Mik15, Kou01, Kou02, Pta08, Log02, Log04) can be seen as a mixture of the two predominant clusters. Furthermore, all individuals except Log04 show a small Natufian related component. Neither of our Early Bronze Age island individuals (Pta08, Kou01 and Kou03) show European HG component. Mik15 seems to have a residual amount of European HG (the bootstrap does not exclude it as this component error bar does not overlap zero, Figure 4.5). Our estimates indicate that Bronze Age Greek samples from Logkas have the higher proportion of the European HG cluster (the main component from steppe populations). These Logkas samples are thus the oldest in Greece showing the influence of a possible steppe-related gene flow. Interestingly, we found higher estimates of steppe-related populations in Logkas (North Greece) than in the Mycenaeans (South Greece) from (Lazaridis et al. 2017). This higher proportion in Logkas is seen for all admixture estimates for $K > 5$, which is consistent with three alternative hypotheses: 1) Logkas had a higher admixture with populations that carried the Steppe component into mainland Greece due to a longer contact or higher gene flow; 2) Logkas had a larger effective size than other Mycenean populations and hence maintained the ancestral diversity; or 3) there was genetic structure within Helladic culture between either north and south Greece or Peloponnese and non-Peloponnese people. Our results do not differ from the results of Lazaridis et al. 2017, who describe samples from Minoans (2400-1700 BCE) and Mycenaeans (1411 –1262) from late Bronze Age as an homogenous group, sharing the 'local' Aegean Farmers component and an 'eastern' Caucasus-like component (Lazaridis et al. 2017). They also detect a 'northern' steppe related ancestry in Mycenaeans, which is recovered in our analyses that included their samples. One question that remained unanswered in Lazaridis et al. 2017 was when did the gene flow events that lead to introduction of 'eastern' and 'northern' component into the Aegean took place. With our data and admixture results we found that the 'eastern' Caucasus/Iran-related component was present across mainland Greece and Aegean islands in the beginning of the Bronze Age. In contrast, the 'northen' European HG, steppe-related ancestry, was mostly found in samples from mainland Greece from middle Bronze Age in Logkas.

We see no substantial difference between Minoans and our Cycladic (Pta08) components. People from Helladic culture show an extra steppe-related ancestry, maxed out in North Greece. Because there is little

difference between Mycenaeans, from the Peloponnese, Minoans and Cycladic individuals, they seem to be genetically homogeneous. This suggests that genetic structure does not reflect cultural divisions between the three main civilizations that emerged in the Aegean during the Bronze Age.

In the middle late bronze age (MLBA), the samples from steppe populations start to show an increase of the component associated with Neolithic Farmers, in agreement with the estimates of (de Barros Damgaard et al. 2018). According to (Lazaridis et al. 2016), the Armenian Chalcolithic samples are the first to derive ancestry from Steppe-related populations in the Caucasus region, giving rise to what then become the typical ancestry of Yamnaya Pastoralists. Our results also show the European HG component appearing in Armenian samples. Yamnaya are thought to be the origin of the Proto-Indo European language (the origin of almost all Eurasian languages spoken nowadays) (de Barros Damgaard et al. 2018). De Barros and colleagues cannot reject that the Proto Indo-European could have evolved under the influence of a Caucasus language (de Barros Damgaard et al. 2018); they point out that this is contrary to previous views that Proto Indo-European language had originated in the steppes north of the Caucasus. Our admixture results show (as the ones from de Barros Damgaard et al. 2018) that Early and Middle Bronze Age Steppe (from which the Yamnaya Karagash belongs - individual of the Stepppe EMBA on both Figure 4.4 and Supplementary Figure 1) have Iranian and Caucasus Hunter-Gatherer components. Thus, we cannot exclude the possibility that the Proto-Indo European language evolved within Caucasus influence. Our Early Middle Bronze Age from the Steppes diverges from the other Bronze Age samples by lacking the Neolithic Farmer component. Interestingly, this component re-appears later on Middle and Late Bronze Age Steppes.

The fact that Bronze Age Anatolians do not show ancestry derived from the Steppes, and instead are formed by the Neolithic Farmer and Caucasus/Iran-like components, suggest limited contact between Anatolia and the Steppe populations, in agreement with the results of (de Barros Damgaard et al. 2018). In Europe, based on the samples from Spain, Germany, Poland and Hungary, we see a clear increase of the European HG (steppe-related) component from the Neolithic (Europe MNChl) to the Late Neolithic and Early Bronze Age (Europe LNBA). This in accordance with (Haak et al. 2015) which plots Germans and Hungarians from Bronze Age between Neolithic people from the same region and Yamnaya.

Modern day Cypriots show almost no European HG component while having more of Iran/Caucasus-like component, which is not surprising as they are geographically closer to the Caucasus. Greeks and Cretans have ancestry proportions very similar to Middle Bronze Age samples from Logkas. We show that the transition from Neolithic to Bronze Age was pivotal to the genetic ancestry of modern-day European populations. In fact, all European populations can be modelled as a mixture of Western Hunter-Gatherers, Neolithic Europeans and Yamnaya, with lower values of Yamnaya related ancestry in southern Europe (Haak et al. 2015). One interpretation for this is a three-wave model where present-day individuals represent the result of migration followed by admixture of Neolithic Europeans, Steppe-related populations and populations from the Caucasus (Lazaridis I, Patterson N, Mittnik A, 2014).

Finally, it is important to stress that Admixture results should be interpreted with caution. Here, I have interpreted changes in admixture proportions from different clusters as evidence of gene flow. However, other demographic events may lead to the same results. In a simulation study, (Lawson, van Dorp, and Falush 2018) showed how three distinct scenarios (recent admixture, recent bottleneck or having an unsampled population, denoted as 'ghost population') can produce the same admixture estimates. For this reason, we can only interpret *ADMIXTURE* estimates as an indication of shared ancestry. The effective size of a given population affects admixture estimates, as larger effective sizes will have less differentiation

meaning that they can have more components than small populations that diverge faster due to genetic drift. Samples with extremely missing data also may result in individuals getting more artifact components in the admixture plot, as there is less information. To account for missing data we also estimated the bootstrap standar error. By comparing the bootstrap standard error for our newly sequenced six samples with the ones from Lazaridis we see that the error intervals for the Neolithic Farmer and Caucasus/Iran-related component is higher for the samples with more missing data from Minoan Odigitria (the ones with ~90% missing data). This raises the question of whether investing more money and having less samples with higher quality is worth over investing the same money for more individuals. Regarding our bootstrap errors alone the missingness of the samples affects the bias of the estimates but we can still detect differences in the proportions of the four major ancestry components.

5.4. Runs of Homozygosity

We estimated Runs of Homozygosity as a proxy for the past effective size of populations. Our starting hypothesis was that Bronze Age Aegean civilizations in islands had a lower effective size than mainland civilization. Estimates of the effective size are also important to distinguish among alternative scenarios that could explain the larger proportion of European HG component in Middle Bronze Age samples from mainland Greece (Logkas), as this could be due to more admixture (or a longer period of gene flow) with populations that carried a steppe component or because Logkas had a higher effective size, maintaining the ancestral diversity for longer than islands sampled. The Runs of Homozygosity could also enlighten if island individuals (Kou01, Kou03 and Pta08) were more inbred than mainland individuals (Log02, Log04 and Mik15). Although we used the same conditions applied by (Ceballos, Hazelhurst, and Ramsay 2018) for detecting ROHs in low-coverage data, we were not able to detect any ROH larger than 0.3Mb using PLINK. We developed a heuristic approach to detect stretches of homozygous sites based on the genotype likelihoods. Due to time constraints, we did not validate the method with simulations or by applying it to analyze already published high quality datasets where ROHs were detected with confidence. Still, we applied the method to the newly sequenced data from the six individuals from Bronze Age. We detected ROHs larger than 0.1MB in chromosome 21, but these should be seen as preliminary data. Our preliminary results indicate that both Bronze Age Greeks from island and mainland are not inbred as the total number of ROH over 0.1 Mb (NROH) was very low on all samples (35 was the most and was found in Kou01) and the total sum of those ROH (SROH) also very low (< 6Mb, also in Kou01) (Figure 4.6). The sample with more NROH and SROH was Kou01 and the one with the lowest was Kou03. The fact that we have such low NROH and SROH suggests low levels of inbreeding in our samples and that both mainland and islands had relatively high effective sizes. This could be caused by the likely increase effective sizes and increase of movement of people within Aegean cultures during the Bronze Age. We only run our approach on chromosome 21 due to time constrains, but it would be interesting to extend this analysis to the other autosomal chromosomes. One reason why no long ROH are detected could be because we have ancient DNA with low coverage. This has two effects: i) due to low coverage most genotypes will have a high uncertainty with relative likelihoods lower than 0.80 for each possible genotype; or ii) the DNA damage patterns introduce biases that affect our ability to detect ROHs, as homozygous sites appear as having a higher probability of being heterozygous. A way of validating and testing this would be to either: i) down sample a genome for which we already know based on other methods that there are long runs of homozygosity and introducing ancient DNA damage pattern at a given error rate; or 2) simulate data with long ROHs and apply our method and Plink to verify under which conditions our approach detects the same ROHs as Plink.

6. Conclusion and Future Work

In this work we were exposed to the particularities of working with ancient genetic material. When dealing with ancient DNA we need to be specially concerned with the quality in terms of base quality, depth of coverage and percentage of the genome covered by sequencing. If necessary, we should: i) remove individuals with high proportion of missing data; ii) remove transitions; and iii) confirm if the results we obtain are not biased by merging modern and ancient samples, oversampled populations or deamination patterns. One way to test this is by comparing analysis done with and without transitions and verifying if, in the case of population structure analysis we have a clear separation between our individuals and an outgroup population.

Our population structure analyses show that in the region surrounding the Aegean, there were changes in the main genetic components during the transition from Neolithic to Bronze Age. This coincided with the appearance of individuals with admixture proportions from mainly three groups: i) Caucasus and Iran, ii) Natufian-related (sampled in todays Israel); iii) European HG. This eastern influence from the Caucasus, Iran and, in lower proportion Israel, increases during the Bronze Age both in mainland and in the islands of the Aegean. This suggests that there was a shift from Greek Neolithic samples where we estimated mostly around 80% proportions attributed to the Neolithic farmers component. The fact that Bronze Age individuals exhibit a mixture of this Neolithic farmer component with Caucasus/Iran-related and Natufian are consistent with gene flow from East into the Aegean region, affecting islands and mainland in similar ways. Thus, this suggests that the Bronze Age transition was associated with the movement of people and gene flow, which could also reflect the likely increase in trade and contact between different populations.

Later in the Bronze Age, we detect an increase of the European Hunter-Gatherer ancestry, related to the steppes. This is found to be higher in samples from Logkas in northern Greece then in the Peloponnese. This is the earliest evidence for the influence of steppe-related populations in Greece, as our Logkas samples predate the other Bronze Age Greek samples published previously. Based on our results, it seems that within Helladic culture people from the Peloponnese have a lower amount of steppe-related ancestry than the samples from the North. This could be because of genetic structure between north and south mainland Greece, with higher admixture with populations with steppe-related ancestry in northern Greek populations, or due to a longer period of contact in the north or higher effective size for the northern populations, meaning that the population with larger effective size would maintain ancestral diversity for longer. Our results also indicate that the cultural divisions from the three Aegean civilizations are not reflected in clear genetic structure as we do not find major differences in the admixture estimates between civilizations. All samples have Anatolian Neolithic and Eastern-related ancestry, with mainland having an extra European Hunter-Gatherer component. Our Middle Bronze Age samples show similar component proportions with Modern Greeks and Cretans, suggesting that movements of people during Bronze Age established what is now the genetic ancestry of present-day Greek people. Sampling more Bronze Age individuals of different ages from the three civilizations would be important to better understand the timing and spread of the introduction of the Eastern and European HG ancestries that we detected. Furthermore, to distinguish among alternative scenarios is would be necessary to perform model comparison or parameter estimation using, for instance, Approximate Bayesian computation (ABC) or other existing methods to infer the demographic history, including past population effective sizes, times of split and migration rates of Bronze Age Greece.

Although we did not detect ROHs with *PLINK*, we detected very small ones (no larger than 0.31 Mb) with our new method. We had more of those small segments in two of our island individuals. Because we did it only on Chromosome 21, it would be interesting to do for the other autosomal chromosomes. Further testing with either simulated data or using a down sampled genome for which we have known ROHs would be needed to check how well it performs compared to *PLINK* and other methods to detect ROHs.

The method for sampling one read is a common practice in ancient DNA studies given the quality and low coverage of some samples. However, it remains unclear to which extent it affects the ancestry estimation. We could test it by having a panel of high-coverage WGS ancient genomes and estimating ancestry proportions and then sampling one read and comparing the estimations afterwards. If research groups invested more money to have higher quality genomes, when the preservation of the DNA allows, it may not be needed to apply this read sampling method. Having higher quality genomes or using genotyple likelihood methods would also allow to more efficiently perform other analysis, such as ROH estimation and model-based inference to reconstruct the demographic history using ABC. Using Genotype likelihoods methods, such as the ones implemented in ANGSD (Korneliussen, Albrechtsen, and Nielsen 2014) and NGSAdmix (Skotte, Korneliussen, and Albrechtsen 2013), reflect the uncertainty on the data and hence provide more information than sampling one read at random. This was what we aimed for in the development of the method to detect ROHs. However, for the study of population structure we did not apply NGSadmix as we were merging WGS data for which we have genotype likelihoods, with a previously published SNP array data for which we did not have access to genotype likelihoods.

Regarding both scripts for sampling one read and detecting ROHs, they are publicly available on github (https://github.com/FCoroado?tab=repositories) and will be further polished, documented and optimized in the future, so that other researchers could use them to analyze genomes from low coverage data.

Bibliography

Alexander, David H, and John Novembre. 2015. “Admixture 1 .3 Software Manual.”

Allentoft, Morten E., Martin Sikora, Karl Göran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B. Damgaard, et al. 2015. “Population Genomics of Bronze Age Eurasia.” *Nature* 522 (7555): 167–72. https://doi.org/10.1038/nature14507.

Barros Damgaard, Peter de, Rui Martiniano, Jack Kamm, J. Víctor Moreno-Mayar, Guus Kroonen, Michaël Peyrot, Gojko Barjamovic, et al. 2018. “The First Horse Herders and the Impact of Early Bronze Age Steppe Expansions into Asia.” *Science* 360 (6396): eaar7711. https://doi.org/10.1126/science.aar7711.

Biehl P., Rassamakin Y. 2008. *Import and Imitation in Archaeology*. Edited by Beier & Beran.

Binladen, Jonas, Carsten Wiuf, M. Thomas P. Gilbert, Michael Bunce, Ross Barnett, Greger Larson, Alex D. Greenwood, et al. 2006. “Assessing the Fidelity of Ancient DNA Sequences Amplified from Nuclear Genes.” *Genetics* 172 (2): 733–41. https://doi.org/10.1534/genetics.105.049718.

Burger, J., M. Kirchner, B. Bramanti, W. Haak, and M. G. Thomas. 2007. “Absence of the Lactase-Persistence-Associated Allele in Early Neolithic Europeans.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (10): 3736–41. https://doi.org/10.1073/pnas.0607187104.

Ceballos, Francisco C., Scott Hazelhurst, and Michèle Ramsay. 2018. “Assessing Runs of Homozygosity: A Comparison of SNP Array and Whole Genome Sequence Low Coverage Data.” *BMC Genomics* 19 (1): 1–12. https://doi.org/10.1186/s12864-018-4489-0.

Ceballos, Francisco C., Peter K. Joshi, David W. Clark, Michèle Ramsay, and James F. Wilson. 2018. “Runs of Homozygosity: Windows into Population History and Trait Architecture.” *Nature Reviews Genetics* 19 (4): 220–34. https://doi.org/10.1038/nrg.2017.109.

Chang, Christopher C., Carson C. Chow, Laurent C.A.M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets.” *GigaScience* 4 (1): 1–16. https://doi.org/10.1186/s13742-015-0047-8.

Copley, M. S., R. Berstan, S. N. Dudd, G. Docherty, A. J. Mukherjee, V. Straker, S. Payne, and R. P. Evershed. 2003. “Direct Chemical Evidence for Widespread Dairying in Prehistoric Britain.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (4): 1524–29. https://doi.org/10.1073/pnas.0335955100.

Dabney, Jesse, Matthias Meyer, and Svante Pääbo. 2013. “Ancient DNA Damage.” *Cold Spring Harbor Perspectives in Biology* 5 (7): 1–7. https://doi.org/10.1101/cshperspect.a012567.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.

Fort, Joaquim. 2015. “Demic and Cultural Diffusion Propagated the Neolithic Transition across Different Regions of Europe.” *Journal of the Royal Society Interface* 12 (106). https://doi.org/10.1098/rsif.2015.0166.

Francis, R. M. 2017. “Pophelper: An R Package and Web App to Analyse and Visualize Population Structure.” *Molecular Ecology Resources* 17 (1): 27–32. https://doi.org/10.1111/1755-0998.12509.

Gaunitz, Charleen, Antoine Fages, Kristian Hanghøj, Anders Albrechtsen, Naveed Khan, Mikkel Schubert, Andaine Seguin-Orlando, et al. 2018. “Ancient Genomes Revisit the Ancestry of Domestic and Przewalski’s Horses.” *Science* 360 (6384): 111–14. https://doi.org/10.1126/science.aao3297.

Gerbault, Pascale, Anke Liebert, Yuval Itan, Adam Powell, Mathias Currat, Joachim Burger, Dallas M. Swallow, and Mark G. Thomas. 2011. “Evolution of Lactase Persistence: An Example of Human Niche Construction.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 366 (1566): 863–77. https://doi.org/10.1098/rstb.2010.0268.

González-Fortes, Gloria, Eppie R. Jones, Emma Lightfoot, Clive Bonsall, Catalin Lazar, Aurora Grandal-d’Anglade, María Dolores Garralda, et al. 2017. “Paleogenomic Evidence for Multi-Generational Mixing between Neolithic Farmers and Mesolithic Hunter-Gatherers in the Lower Danube Basin.” *Current Biology* 27 (12): 1801-1810.e10. https://doi.org/10.1016/j.cub.2017.05.023.

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, et al. 2010. “A Draft Sequence of the Neandertal Genome.” *Science* 328 (5979): 710–22. https://doi.org/10.1126/science.1188021.

Haak, Wolfgang, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, et al. 2015. “Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe.” *Nature* 522 (7555): 207–11. https://doi.org/10.1038/nature14317.

Hofmanová, Zuzana, Susanne Kreutzer, Garrett Hellenthal, Christian Sell, Yoan Diekmann, David Díez-Del-Molino, Lucy Van Dorp, et al. 2016. “Early Farmers from across Europe Directly Descended from Neolithic Aegeans.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (25): 6886–91. https://doi.org/10.1073/pnas.1523951113.

J. and Lange, K. Alexander D.H. Novembre. 2009. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” *Genome Research* 19 (9): 1655–64. https://doi.org/10.1101/gr.094052.109.

Jones, Eppie R., Gloria Gonzalez-Fortes, Sarah Connell, Veronika Siska, Anders Eriksson, Rui Martiniano, Russell L. McLaughlin, et al. 2015. “Upper Palaeolithic Genomes Reveal Deep Roots of Modern Eurasians.” *Nature Communications* 6: 1–8. https://doi.org/10.1038/ncomms9912.

Juras, Anna, Maciej Chyleński, Edvard Ehler, Helena Malmström, Danuta Żurkiewicz, Piotr Włodarczak, Stanisław Wilk, et al. 2018. “Mitochondrial Genomes Reveal an East to West

Cline of Steppe Ancestry in Corded Ware Populations." *Scientific Reports* 8 (1): 1–10. https://doi.org/10.1038/s41598-018-29914-5.

Kircher, Martin, Susanna Sawyer, and Matthias Meyer. 2012. "Double Indexing Overcomes Inaccuracies in Multiplex Sequencing on the Illumina Platform." *Nucleic Acids Research* 40 (1): 1–8. https://doi.org/10.1093/nar/gkr771.

Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen. 2014. "ANGSD: Analysis of Next Generation Sequencing Data." *BMC Bioinformatics* 15 (1). https://doi.org/10.1186/s12859-014-0356-4.

Lawson, Daniel J., Lucy van Dorp, and Daniel Falush. 2018. "A Tutorial on How Not to Over-Interpret STRUCTURE and ADMIXTURE Bar Plots." *Nature Communications* 9 (1): 1–11. https://doi.org/10.1038/s41467-018-05257-7.

Lazaridis, Iosif, Alissa Mittnik, Nick Patterson, Swapan Mallick, Nadin Rohland, Saskia Pfrengle, Anja Furtwängler, et al. 2017. "Genetic Origins of the Minoans and Mycenaeans." *Nature* 548 (7666): 214–18. https://doi.org/10.1038/nature23310.

Lazaridis, Iosif, Dani Nadel, Gary Rollefson, Deborah C. Merrett, Nadin Rohland, Swapan Mallick, Daniel Fernandes, et al. 2016. "Genomic Insights into the Origin of Farming in the Ancient Near East." *Nature* 536 (7617): 419–24. https://doi.org/10.1038/nature19310.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, et al. 2016. "The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations." *Nature* 538 (7624): 201–6. https://doi.org/10.1038/nature18964.

Mathieson, Sara, and Iain Mathieson. 2018. "FADS1 and the Timing of Human Adaptation to Agriculture." *Molecular Biology and Evolution* 35 (12): 2957–70. https://doi.org/10.1093/molbev/msy180.

McQuillan, Ruth, Anne Louise Leutenegger, Rehab Abdel-Rahman, Christopher S. Franklin, Marijana Pericic, Lovorka Barac-Lauc, Nina Smolej-Narancic, et al. 2008. "Runs of Homozygosity in European Populations." *American Journal of Human Genetics* 83 (3): 359–72. https://doi.org/10.1016/j.ajhg.2008.08.007.

Mittnik A, et al. Lazaridis I Patterson N. 2014. "Ancient Human Genomes Suggest Three Ancestral Populations for Present-Day Europeans." *Nature* 513 (7518): 409–13. https://doi.org/10.1038/nature13673.

Narasimhan, Vagheesh, Petr Danecek, Aylwyn Scally, Yali Xue, Chris Tyler-Smith, and Richard Durbin. 2016. "BCFtools/RoH: A Hidden Markov Model Approach for Detecting Autozygosity from next-Generation Sequencing Data." *Bioinformatics* 32 (11): 1749–51. https://doi.org/10.1093/bioinformatics/btw044.

Orlando, Ludovic. 2019. "Ancient Genomes Reveal Unexpected Horse Domestication and

Management Dynamics." *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* 1900164: e1900164. https://doi.org/10.1002/bies.201900164.

Parang, Keykavous, Leonard Wiebe, and Edward Knaus. 2012. *Novel Approaches for Designing 5-O-Ester Prodrugs of 3-Azido-2,3-Dideoxythymidine (AZT). Current Medicinal Chemistry*. Vol. 7. https://doi.org/10.2174/0929867003374372.

Patterson, Nick, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. 2012. "Ancient Admixture in Human History." *Genetics* 192 (3): 1065–93. https://doi.org/10.1534/genetics.112.145037.

Patterson, Nick, Alkes L. Price, and David Reich. 2006. "Population Structure and Eigenanalysis." *PLoS Genetics* 2 (12): 2074–93. https://doi.org/10.1371/journal.pgen.0020190.

Pinhasi, Ron, Daniel Fernandes, Kendra Sirak, Mario Novak, Sarah Connell, Songül Alpaslan-Roodenberg, Fokke Gerritsen, et al. 2015. "Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone." *PLoS ONE* 10 (6): 1–13. https://doi.org/10.1371/journal.pone.0129102.

Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38 (8): 904–9. https://doi.org/10.1038/ng1847.

Raj, Anil, Matthew Stephens, and Jonathan K Pritchard. 2013. "Variational Inference of Population Structure in Large SNP Datasets Corresponding Author :" *BioRxiv*, 1–40. https://doi.org/10.1534/genetics.114.164350.

Reich, David, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, et al. 2010. "Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia." *Nature* 468 (7327): 1053–60. https://doi.org/10.1038/nature09710.

Schofield, Elizabeth, and Cyprian Broodbank. 2002. *An Island Archaeology of the Early Cyclades*. *American Journal of Archaeology*. Vol. 106. Cambridge University Press. https://doi.org/10.2307/4126226.

Sikora, Martin, Vladimir V. Pitulko, Vitor C. Sousa, Morten E. Allentoft, Lasse Vinner, Simon Rasmussen, Ashot Margaryan, et al. 2019. "The Population History of Northeastern Siberia since the Pleistocene." *Nature*. https://doi.org/10.1038/s41586-019-1279-z.

Skoglund, Pontus, Helena Malmström, Maanasa Raghavan, Jan Storå, Per Hall, Eske Willerslev, M. Thomas P Gilbert, Anders Götherström, and Mattias Jakobsson. 2012. "Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe." *Science* 336 (6080): 466–69. https://doi.org/10.1126/science.1216304.

Skoglund, Pontus, and Iain Mathieson. 2018. "Ancient Genomics of Modern Humans: The First Decade." *Annual Review of Genomics and Human Genetics* 19 (1): 381–404. https://doi.org/10.1146/annurev-genom-083117-021749.

Skotte, Line, Thorfinn Sand Korneliussen, and Anders Albrechtsen. 2013. “Estimating Individual Admixture Proportions from next Generation Sequencing Data.” *Genetics* 195 (3): 693–702. https://doi.org/10.1534/genetics.113.154138.

Tang, Hua, Jie Peng, Pei Wang, and Neil J. Risch. 2005. “Estimation of Individual Admixture: Analytical and Study Design Considerations.” *Genetic Epidemiology* 28 (4): 289–301. https://doi.org/10.1002/gepi.20064.

Telenti, Amalio, Levi C.T. Pierce, William H. Biggs, Julia Di Iulio, Emily H.M. Wong, Martin M. Fabani, Ewen F. Kirkness, et al. 2016. “Deep Sequencing of 10,000 Human Genomes.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (42): 11901–6. https://doi.org/10.1073/pnas.1613365113.

Willerslev, Eske, and Alan Cooper. 2005. “Review Paper. Ancient DNA.” *Proceedings of the Royal Society B: Biological Sciences* 272 (1558): 3–16. https://doi.org/10.1098/rspb.2004.2813.

Supplementary Material

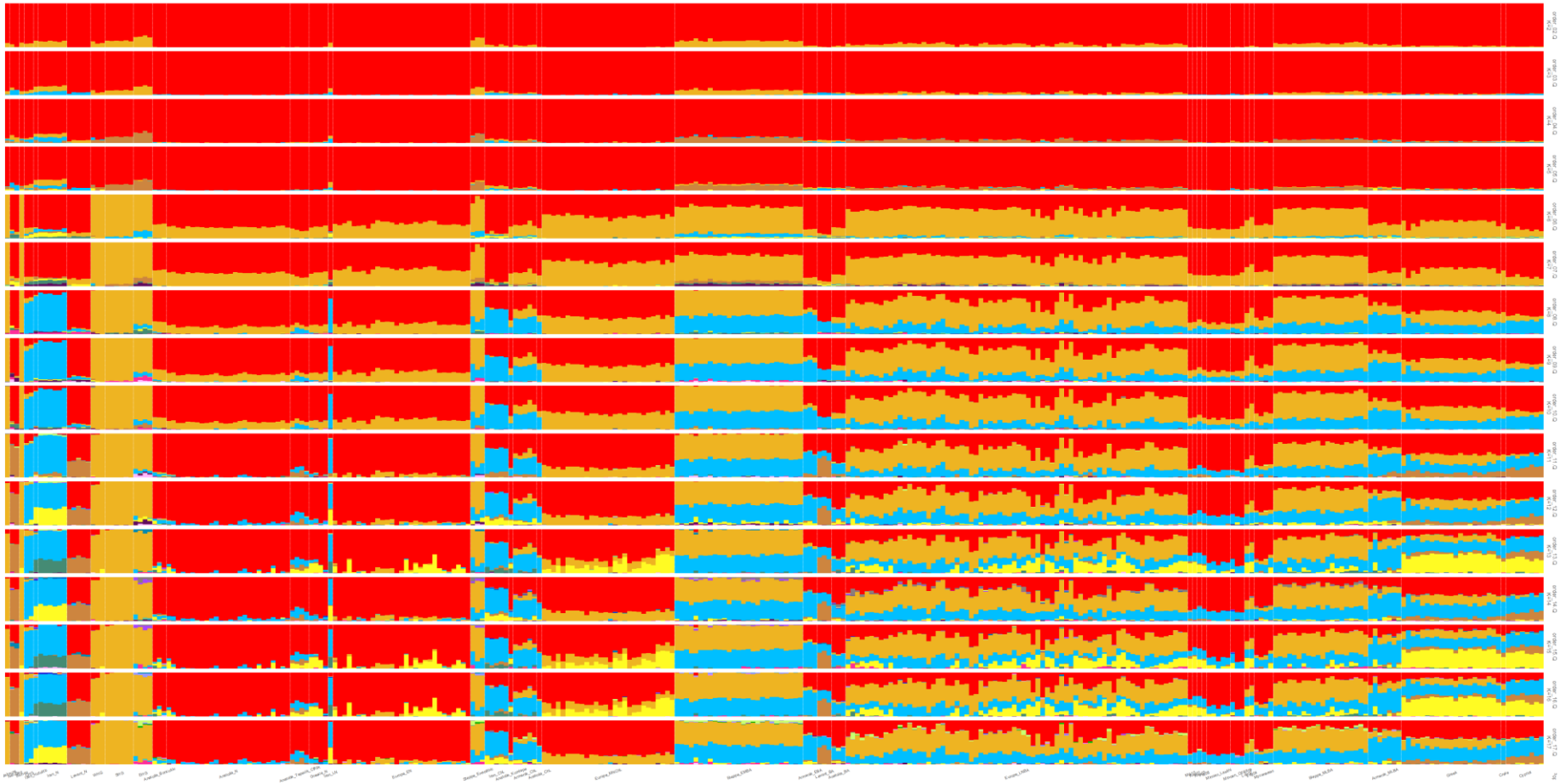

Figure SI1 Admixture plot for all K (ranging from 2 to 17), for all relevant population ordered by age.

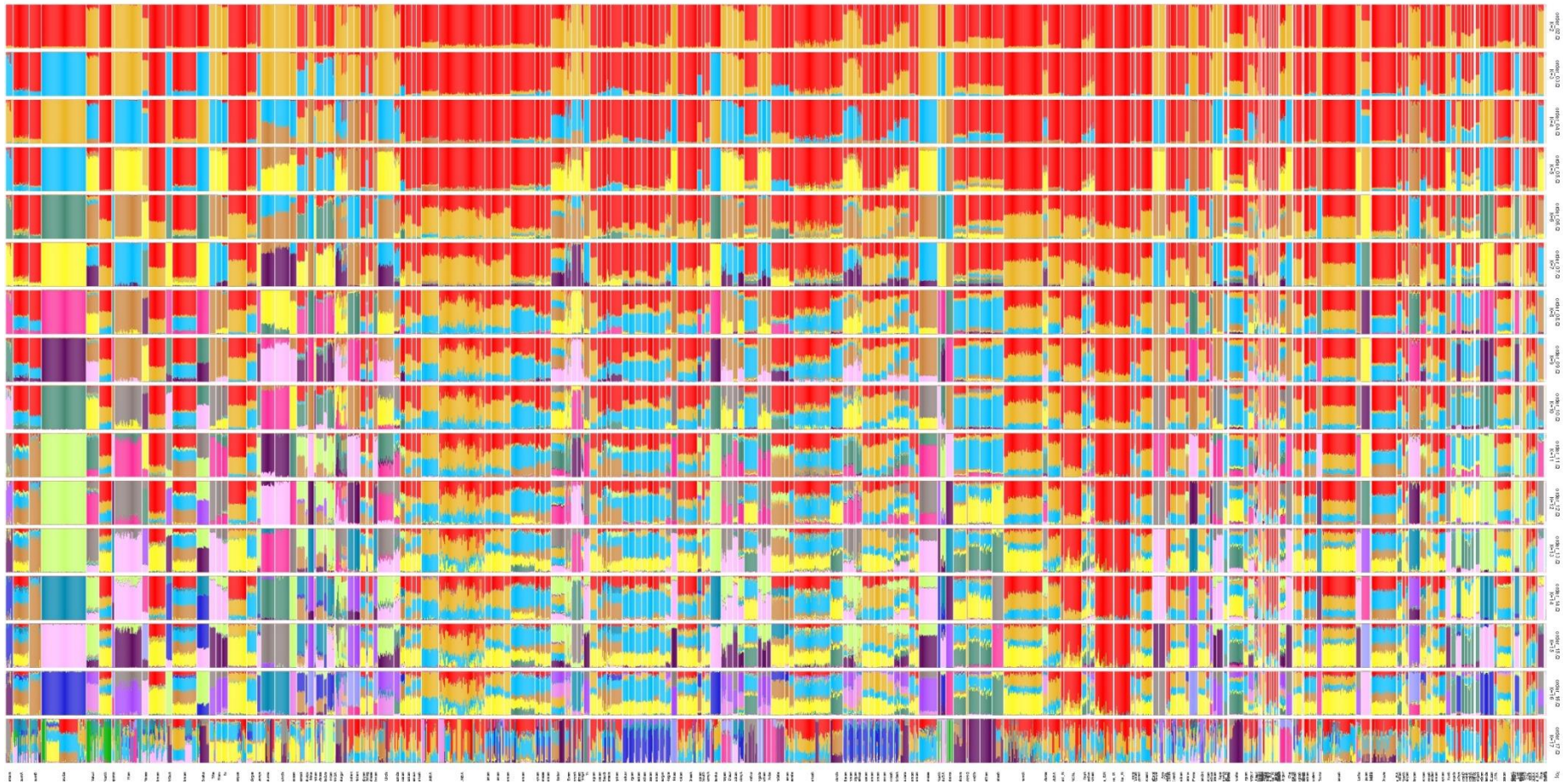

Figure SI2- Admixture plot for all K (ranging from 2 to 17), for all 2399 Individuals used to run admixture